Effective Classification after Dimension Reduction: A Comparative Study

Mohini D Patil[*], Dr. Shirish S. Sane

[*] PG Student, Department of Computer Engineering, K.K.W.I.E.E.R, Pune University.
[**] Head of Department of Computer Engineering, K.K.W.I.E.E.R, Pune University.

Abstract- Classification is undoubtedly gaining major importance in the fields of machine learning, pattern recognition genetic engineering and bio medical sciences where it can be used for automated decision making. Mostly these areas contain datasets having large number of dimensions which require some preprocessing. Thus dimension reduction is a preprocessing step carried out prior to classification so that the classifiers can be designed in an easy to compute way. However while doing so, it must also retain the accuracy and must not lead to loss of information. Thus effective classification must be carried out by employing proper dimension reduction techniques. The paper discusses in brief about the dimension reduction techniques. It also describes the system developed for dimension reduction and use of the tool WEKA for dimension reduction and preprocessing. Finally a comparative study of the results obtained by the system and WEKA is done.

Index Terms- Classification, Dimension Reduction, Preprocessing, WEKA

I. INTRODUCTION

Reduction of dimensions(attributes) of large datasets has always been an area of research, specially for the datasets involved in the field of medical science, genetics and bio engineering. These datasets have dimensions of the order of thousands and not all of them may be relevant for classification purpose. From the point of view of classification applications, it is important to retain only those attributes which help to increase the effectiveness. This also helps in designing simple and easy to compute algorithms. Dimension reduction covers not only the attribute reduction but also the instance reduction. But the major focus has always been on attribute reduction. Thus when preprocessing of the data is done for dimension reduction, it must focus on describing the dataset using minimum number of attributes but it must give the performance comparable to that of original dataset containing all the attributes. Basically two categories of solutions have been described in the literature [1][2] for attribute reduction. The first one is attribute selection[1] which chooses the attributes to be retained in the reduced dataset and removal of the remaining. Second approach to reduce the dimensions is by attribute extractions[1] where the data contained in the original attributes can be completely used up and new dimensions can be generated which are richer in content as compared to original. Each of these methods has their own pros and cons. A midway between the two is combining the merits of both where a system can either perform selection or extraction depending upon the problem to be solved. A brief summary of this has been given in our earlier work [3] . Also there are different criteria and factors which need to be considered while choosing the dimensions for retention like relevance, significance, dependency etc. And these criteria are based on theories like rough sets, fuzzy sets, and fuzzy rough sets [4]. Once the dataset is reduced it needs to be combined with a classifier to check for accuracy and the best combination needs to be determined. Different classifiers like nearest neighbor and decision trees are already in the field for better classification. Besides them, neural network classifiers have been an emerging classifier these days which is hoped to have a better performance. Thus if a system can be designed which can reduce the dimensions and perform classification with neural network classifiers, it could provide an effective solution for many applications in the field of medical science. This has been described in the paper. Also WEKA [5] has been a very powerful tool designed for machine learning and data mining. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It provides various algorithms for data preprocessing which includes the ones for attribute selection and extraction. The paper gives a description about the background of dimension reduction; system designed; the inbuilt algorithms used by WEKA for dimension reduction and finally tabulates the comparison of the results.

II. TECHNIQUES FOR DIMENSION REDUCTION

This section describes some of the techniques available in literature for dimension reduction.

- In [6], Ron.Kohavi et.al has described various wrapper techniques for selecting the features. It mentions about the FOCUS algorithm which makes use of exhaustive technique to try out all the subset of features and selects the minimum subset which can be sufficient to determine the class value for all the objects of training set. It also describes RELIEF algorithm which makes use of the relevance criteria to determine the importance of the features. The relevance values are computed by finding the difference values between the selected instance and two nearest instances of same and opposite class. Thus it attempts to find all the relevant features.
- In [7], Isabelle Guyon et.al have described about the SVM-RFE(Support Vector machine- Recursive Feature Elimination) algorithm which makes use of the weight criteria for ranking the features. It first selects training examples having good feature indices and then trains the SVM classifier. Then it computes the weight vector

and finds the ranking criteria. Feature that has the smallest ranking criterion is removed from the list and further iterations are carried out.

- In [8] , Yijun Sun et.al have proposed an algorithm for feature selection in microarray datasets. It first finds out the margin for each sample. For this it makes use of the neighbors of the sample, one from same class and other from different class. From them, the margin is computed by using Manhattan/ Euclidian distance. This computation gives an idea about how many features of the object may be corrupted by noise and helps to minimize the induced feature space. Thus the algorithm transforms arbitrary non linear problem into set of locally linear ones. This is then used to estimate the feature weights. A probabilistic model has been employed for computation.
- In[9],Hua-Liang Wei et.al have described the FOS(Forward Orthogonal Search) algorithm. It makes use of the squared correlation function as the criteria for measuring the dependency between the features and keeps on selecting the features in a stepwise way. In every iteration, such feature is selected by which the resultant candidate subset can represent the original dataset containing all the features.
- In [4],R.Jensen et.al have discussed about the Quick Reduct and Fuzzy Rough Quick Reduct algorithms. The algorithms work on the same principle, they keep on adding one by one the attribute which increases the dependency with respect to the decision label and stop when the dependency no longer increases i.e when the dependency of the reduced set equals the dependency of original attribute set. The only difference is in the criteria used for finding the dependency, the Quick Reduct algorithm makes use of the measures based on rough set theory while that of Fuzzy Quick Reduct makes use of measures based on the fuzzy rough theory.

III. DIMENSION REDUCTION IN WEKA

WEKA is a software workbench developed for supporting machine learning and supports a number of activities including the preprocessing activities, classification, clustering and visualization [5]. As a part of the preprocessing step, the dimensions of large datasets can be reduced in WEKA by following ways

A. Supervised Filter Approach

WEKA provides a supervised attribute filter that can be flexible and allows various search and evaluation methods to be combined. It makes use of two things that are evaluator and search. Evaluator determines the criteria used for selecting the attribute.

Different evaluators [10] that can be used are:

- CfsSubsetEval: It considers the predictive ability of each feature and finds out the relevance. It also checks for redundancy between the selected features. Finally the subset of attributes which are highly coorelated with the class and are less redundant is chosen.
- CorrelationAttributeEval: It evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average.
- GainRatioAttributeEval: The gain ratio with respect to class label is computed and this determines the worth of attribute for selection.
- OneRAttributeEval: Evaluates the worth of an attribute by using the OneR classifier.
- ReliefAttributeEval: It is based on the RELIEF algorithm which evaluates the relevance of attribute by considering the difference between the instance and the two nearest instances from same and opposite class. It can operate on both discrete and continuous class data.
- SymmetricalUncertAttributeEval: Evaluates the worth of an attribute by measuring the symmetrical uncertainty with respect to the class.
- WrapperSubsetEval: Evaluates attribute sets by using a learning scheme. Cross validation is used to estimate the accuracy of the learning scheme for a set of attributes.

The search option is provided for determining the way of searching the attribute. Different methods provided[10] are

- BestFirst: It searches for the attribute subset by greedy hill climbing method in combination with backtracking. The backtracking is based on the concept that if some number of consecutive nodes is found such that they do not improve the performance then backtracking is done. It may apply forward approach where it starts from empty set of attributes and goes on adding the next. It may also go for backward approach where it starts from a set of all attributes and removes one by one. It may also adopt a midway between both approaches where search is done in both directions (by considering all possible single attribute additions and deletions at a given point) which is also called as hybrid approach.
- GreedyStepwise: Performs a greedy forward or backward search through the space of attribute subsets. May start with no/all attributes or from an arbitrary point in the space. Stops when the addition/deletion of any remaining attributes results in a decrease in evaluation. Can also produce a ranked list of attributes by traversing the space from one side to the other and recording the order that attributes are selected.
- Ranker: Individual evaluations of the attributes are done and they are ranked accordingly. It is normally used in conjunction with attribute evaluators (Relief, GainRatio, Entropy etc).

B. Unupervised Filter Approach

WEKA provides an unsupervised attribute filter approach namely PCA[10] which is used for extracting the features and reducing the dimensions.

- PrincipalComponents (attribute transformer): Performs a principal components analysis and transformation of the data. Used in conjunction with a Ranker search.

Dimensionality reduction is accomplished by choosing enough eigenvectors to account for some percentage of the variance in the original data---default 0.95 (95%). Attribute noise can be filtered by transforming to the PC space, eliminating some of the worst eigenvectors, and then transforming back to the original space.

IV. PROPOSED SYSTEM

This section briefs about the proposed system.

The steps of working are as:

1. Select the desired dataset
2. Set the number of instances and attributes
3. Check for memory requirement and perform instance selection accordingly.
4. Compute the relevance value for each attribute.
5. Select the attribute having highest relevance
6. Using the selected attribute partition the attributes into subsets containing significant, insignificant and intermediate attributes
7. From the intermediate set select or extract new attribute depending upon the problem at hand
8. Remove the used attributes from the intermediate set and the insignificant attributes
9. If the set of attributes is empty or desired attributes are fetched goto step 11
10. Select next attribute as the one which is relevant and significant and goto step 6
11. Store the resultant dataset containing some selected features / extracted features in WEKA compatible format(.arff)
12. Give the reduced dataset as input to WEKA and test for classification accuracy using training testing method with classifiers like J48(decision trees),KNN(K-Nearest Neighbor) and ANN(Artificial Neural Networks)

The relevance and significance of the attributes is computed using fuzzy rough theory [4][11][12]

V. RESULTS AND DISCUSSIONS

This section describes the results obtained and the comparisons made.

The system was implemented and tested with six datasets namely colon tumor, lung cancer, leukemia, breast cancer, lymphoma and central nervous. The datasets were also given to WEKA and preprocessed using the attribute selection preprocessing filter. The outputs generated by both were to given to classifiers like decision trees (J48),KNN(K nearest neighbor with k=3) and ANN(Artificial Neural Networks) and the accuracies were measured.

The summarized results are tabulated in Table 1. It can be seen that on an average the proposed system gives better accuracy as compared to that of inbuilt filters provided by WEKA. Also for the breast cancer dataset , WEKA could not process the dataset and provided error as insufficient memory heap.

Table 1:Results and Comparison

Data set	No. of Attributes	No. of Records	No. of attributes after reduction	Accuracy after Dimension Reduction in Weka					
				Using Inbuilt Weka Filters			Using Fuzzy Rough Approach		
				J48	ANN	KNN	J48	ANN	KNN
Colon	2001	62	27	87.09	82.25	83.87	72.58	83.87	85.48
Leukemia	7130	34	23	91.11	100	100	97.05	94.11	97.05
Lung	12533	32	41	75	100	100	96.87	100	100
Lymphoma	4027	45	75	82.22	100	100	95.56	97.78	93.33
Central Nervous	7129	60	40	65	80	78.33	100	100	100
Breast Cancer	24482	48	Err	-	-	-	85.41	93.75	93.71
Average				80.08	92.45	92.44	92.39	95.15	95.17

VI. CONCLUSION

We have presented a comparative study on dimension reduction. Firstly we discussed the concept of dimension reduction, its need and areas of application. Then we focused upon some of the techniques used for reducing dimensions. A brief discussion on the inbuilt filters of WEKA for dimension reduction was given. The system was then described and finally we discussed about the results which show that the proposed system can be used as a future direction of computing specially for medical and bio engineering field which has large dimensional datasets. And this is possible because of the fuzzy rough theory which helps to increase the accuracy and also the use of neural network classifiers which provide a good performance. In future, we plan to merge genetic algorithm based dimension reduction with the combination of neural network classifier which may provide comparable or better results in the domain.

REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Technique"s, 2nd ed,The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6.

[2] www.nptel.ac.in

[3] Mohini D Patil and Shirish S Sane,Dimension Reduction: A Review. *International Journal of Computer Applications*92(16):23-29, April 2014. Published by Foundation of Computer Science, New York, USA.

[4] R. Jensen and Q. Shen, "Semantics-preserving dimensionality reduction:Rough and fuzzy rough-based approach" ,IEEE Trans. Knowl. Data Eng., vol. 16, no. 12, pp.14571471, Dec. 2004.

[5] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[6] Ron Kohavi , George H. John , "Wrappers for feature subset selection" ,ELSEVIER, Artificial Intelligence 97 (1997) 273-324

[7] Isabelle Guyon,Jason Weston,Stephen BarnHill, Gene Selection for Cancer Classification using Support Vector Machines, ACM Journal,Machine Learning, Volume 46 Issue 1-3, 2002, Pages 389-422

[8] Yijun Sun, Sinisa Todorovic, and Steve Goodison,LocalLearning Based Feature Selection for High Dimensional Data Analysis,IEEE transactions on Pattern Analysis and Machine Intelligence, Volume 32 issue 9, 1610 - 1626

[9] Hua-Liang Wei and Stephen A. Billings. Feature Subset Selection and Ranking for data dimensionality reduction, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 29, NO. 1, JANUARY 2007

[10] http://weka.sourceforge.net

[11] P.Maji and P.Garai,Fuzzy Rough Simultaneous Attribute Selection and Feature Extraction Algorithm,IEEE Transactions on Cybernetics, VOL. 43, NO. 4,AUGUST 2013

AUTHORS

First Author – Ms.Mohini D Patil, Post Graduate Student, K. K. Wagh Institute of Engineering Education and Research, University of Pune, mohini186@gmail.com

Second Author – Dr. Shirish S. Sane,Head of Computer Department, K. K. Wagh Institute of Engineering Education and Research, University of Pune, ssane@kkwagh.edu.in