Journal of Advanced Computing and Communication Technologies ***(ISSN:* 2347 - 2804*)***
Volume No.1 Issue No. 1, August 2013

A Novel Algorithm for Design Tree Classification with PCA

By

Ravindra Gupta, Gajendra Singh, Gaurav Kumar Saxena

SSSIST Sehore, India

sravindra_p84@rediffmail.com,gajendrasingh86@rediffmail.com,gaurav.saxena18@rediffmail.com

ABSTRACT

Classification technique is useful to categorize data into classes from the current dataset on the basis of a training set of data containing observations (or instances) whose category membership is known. The decision Tree classification algorithm doesn't work well separately for high dimensional data. To improve the efficiency, in our work, we apply principal component analysis and linear transformation together on the original data set for dimensionality reduction after that classification algorithm is applied for a better solution.

Keywords

Classification, Dimensionally Reduction, Principal Component Analysis, Decision Tree.

1. INTRODUCTION

Classification technique is used in the assignment of some combination of input variables, which are measured or preset, into predefined classes. Over the last decade, various technologies came such as imaging processing, gene micro array studies and textual data analysis with huge amount of data or growth of data dimension which may suffer from efficiency of classification rate.

Principal component analysis is appropriate when you have obtained measures on a number of observed variables and wish to develop a smaller number of artificial variables (called principal components) that will account for most of the variance in the observed variables. The principal components may then be used as a predictor or criterion variables in subsequent analyses.

Classification

Classification is the process of generalizing the data according to different instances, in other words we can say that Classification is the task of assigning objects to one of several predefined categories.It is a persistent problem that encompasses many diverse applications such as detecting spam email messages based upon the message header and content, categorizing cells as malignant or benign based upon the results of MRI scans, and classifying galaxies based upon their shapes.

A classification model is useful to distinguish between objects of different classes and also to predict the class label of unknown records. Various methods of classification are decision tree classifiers, rule-based classifiers, neural networks, support vector machines, and naıve Bayes classifiers.

Dimensionality Reduction

In the growing age where various applications are facilitated for us, have a large number of datasets in multidimensional space.Dimensionality reduction deals with the transformation of a high dimensional dataset into a low dimensional space, while retaining most of the useful structure in the original data.For improving the efficiency of classifier dimension reduction plays an important role.

Dimensionality reduction can be done by linear and nonlinear methods, which are described as:

Linear technique is an unsupervised technique, its purpose to perform dimensionality reduction by embedding the data into a subspace of lower dimensionality. Although there exist various techniques to do so, PCA is by far the most popular linear technique.[1]

In mathematical terms, PCA attempts to find a linear mapping M that maximizes M^T cov(X)M, where cov(X) is the covariance matrix of the data X. It can be shown that this linear mapping is formed by the d principal eigenvectors (i.e., principal components) of the covariance matrix of the zero-mean data. Hence, PCA solves the eigen problem

$$\text{cov}(X)M = \lambda M$$

The eigen problem is solved for the d principal eigenvalues λ. The low-dimensional data representations y_i of the datapoints x_i are computed by mapping them onto the linear basis M, i.e., $Y = (X - \bar{X})M$.

PCA has been successfully applied in a large number of domains such as face recognition [2], coin classification [3], and seismic series analysis [4].

The main drawback of PCA is that the size of the covariance matrix is proportional to the dimensionality of the data points. As a result, the computation of the eigenvectors might be infeasible for very high-dimensional data.

On the other hand, Nonlinear techniques for dimensionality reduction must have global and local properties of the original data in the low-dimensional representation and it must perform global alignment of a mixture of linear models.

Principal Component Analysis

Principal Component Analysis (PCA) also known as the Karhunen-Loeve Transform is a classical statistical method. It identifies the axes for a set of data vectors along which the correlation between components of the data vectors can be most clearly shown[43]

Suppose there is a data set M={ Xi | i=1, …, N }, where X is an n-dimensional column vector and X = (x1, … , xn)T. The mean of the data vector is μ = < X >, here < > stands for the average over the data set. The data set can be represented by a matrix D = (X1, X2, …, XN). The covariance matrix of D is C with its element Cij which can be calculated as shown below.

$$Cij = <(xi – \mu i)(xj – \mu j)> \quad \ldots\ldots\ldots\ldots \quad (1.1)$$

By solving the characteristic equation of the covariance matrix C, we can obtain the eigenvectors that specify the axes having the properties described above and the corresponding eigenvalues that are respectively indicative of the variance of the dataset along these axes. Therefore, just by looking at the eigenvalues, we can easily find out along which axes the dataset has little or no spread. Hence, the principal axes and the eigenvalues give a good reflection of the linear interdependence between the components of the data vectors. After choosing some eigenvectors with largest eigenvalues, form a subspace A in which the data set has the most significant amounts of variance. Thus, the dimensionality of the data can be reduced by means of this property of PCA.

Suppose B has all the eigenvectors of the covariance matrix C as its row vectors, we can transform a data vector X this way:

$$Y = B(X- \mu) \ldots\ldots\ldots.. \quad (1.2)$$

By applying this projection to the original dataset D, we can get an uncorrelated vector set {Y}. Since B is an orthogonal matrix, the inverse of B is equal to the transpose of B (BT). We can use Y to obtain the original data vector X like this:

$$X = BTY + \mu \quad \ldots\ldots\ldots.. \quad (1.3)$$

In the subspace A consisting of the eigenvectors having the largest eigenvalues, we can do the similar transformation to the above to get the low dimensional code vector Y' of X.

$$Y' = A(X- \mu) \ldots\ldots\ldots\ldots \quad (1.4)$$

And we can reconstruct X in the way that is similar to (1.3).

$$X' = ATY + \mu \quad \ldots\ldots\ldots\ldots \quad (1.5)$$

By (1.4) and (1.5), we project the original data vector to the low dimensional space spanned by A and then we use the low dimensional code to reconstruct the original data. This projection minimizes the mean-square error between the data and the reconstruction of the data.

In practice, dimensionality reduction using PCA can be done efficiently through singular value decomposition (SVD) [43].

2. RELATED WORK

Dimensionality reduction techniques can be broadly divided into several categories: (i) feature selection and feature weighting, (ii) feature extraction, and (iii) feature grouping.

Dimension reduction techniques based on Feature Selection and Feature Weighting

Feature selection or subset selection deals with the selection of a subset of features that is most appropriate for the task at hand after that feature weighting [5] assigns (usually between zero and one) two different features to indicate the salience of the individual features. Most of the literature on feature selection/weighting pertains to supervised learning.

Feature selection and weighting algorithms categorized on the basis of **Filters, Wrappers, and Embedded** [6].

The filter approaches evaluate the relevance of each feature subset using the data set alone by some learning task. RELIEF and its enhancement are representatives of this class; the basic idea behind this task is to assign feature weights based on the consistency of the feature value in the k nearest neighbors of every data point. Wrapper algorithms,learning algorithm evaluates the quality of each feature (subset Specifically, a learning algorithm (e.g., a nearest neighbor classifier, a decision tree, a naive Bayes method) is run using a feature subset and the feature subset is assessed by some estimate related to the classification accuracy. Often the learning algorithm is regarded as a "black box" in the sense that the wrapper algorithm operates independent of the internal mechanism of the classifier. An example is [9], which used genetic search to adjust the feature weights for the best performance of the k nearest neighbor classifier. In the third approach (called embedded in [10]), the learning algorithm is modified to have the ability to perform feature selection. There is no longer an explicit feature selection step; the algorithm automatically builds a classifier with a small number of features. LASSO (least absolute shrinkage and selection operator) [11] is a good example in this category.

LASSO modifies the ordinary least square by including a constraint on the L1 norm of the weight coefficients. This has the effect of preferring sparse regression coefficients (a formal statement for this is proved in [12,13]), effectively performing feature selection. Another example is MARS (multivariate adaptive regression splines) [13], where choosing the variables used in the polynomial splines effectively performs variable selection. Automatic relevance detection in neural networks [14] is another example, which uses a Bayesian approach to estimate the weights in the neural network as well as the relevancy parameters that can be interpreted as feature weights.

Filter approaches are generally faster because they are classifier independent and require computation of simple quantities. They scale well with the number of features, and many of them can comfortably handle thousands of features. Wrapper approaches, on the other hand, can be superior in accuracy when compared with filters, which ignore the properties of the learning task at hand [15]. They are, however, computationally more demanding, and do not scale very well with the number of features. It is because training and evaluating a classifier with many features can be slow, and the performance of a traditional classifier with a large number of features may not be reliable enough to estimate the utilities of individual features. To get the best results from filters and wrappers, the user can apply a filter-type technique as preprocessing to cut down the feature set to a moderate size, and then use a wrapper algorithm to determine a small yet discriminative feature subset. Some state-of-the-art feature selection algorithms indeed adopt this approach, as observed in. “Embedded" algorithms are highly specialized and it is difficult to compare them in general with filter and wrapper approaches.

Quality of a Feature Subset Feature selection/weighting algorithms can also be classified according to the definition of “relevance" or how the quality of a feature subset is assessed. Five definitions of relevance are given in. Information-theoretic methods are often used to evaluate features, because the mutual information between a relevant feature and the class labels should be high [15]. Non-parametric methods can be used to estimate the probability density function of a continuous feature, which is used to compute the mutual information. Correlation is also used frequently to evaluate features. A feature can be declared irrelevant if it is conditionally independent of the class labels given other features. The concept of Markov blanket is used to formalize this notion of irrelevancy in. RELIEF uses the consistency of the feature value in the k nearest neighbors of every data point to quantify the usefulness of a feature.

Optimization Strategy Given a definition of feature relevancy, a feature selection algorithm can search for the most relevant feature subset. Because of the lack of monotonicity (with relation to the features) of the many feature connexion criteria, a combinatorial search through the area of all doable feature subsets is required. Usually, heuristic (non-exhaustive) methods have to be adopted, because the size of this space is exponential in the number of features. In this case, one generally loses any guarantee of optimality of the selected feature subset. In the last few years different types of heuristics like sequential forward or backward searches, floating search, beam search, bi-directional search, and genetic search have been proposed [16]. A comparison of some of these search heuristics can be found in [17]. In the context of regression toward the mean, successive forward search is commonly referred to as stepwise regression.
Forward stepwise regression is a generalization of stepwise regression, where a feature is only “partially" selected by increasing the corresponding regression coefficient by a fixed amount. It is closely related to LASSO [18], and this relationship was established via least angle regression (LARS), another interesting algorithm on its own, in [20].

Wrapper algorithms have some intelligence with heuristic search. Feature weighting algorithms do not involve a heuristic search because the weights for all features are computed simultaneously. Embedded approaches also do not require any heuristic search. The best parameter is usually calculable by optimizing an explicit objective perform. Counting on the shape of the target perform, completely different optimization methods are used. Within the case of LASSO, as an example, a general quadratic programming problem solver, homotopy methodology [12], a changed version of LARS or the EM algorithmic rule is wanting to estimate the parameters.

Feature Extraction

In feature extraction, a small set of new features is constructed by a general mapping from the high dimensional data. The mapping often involves all the available features. Many techniques for feature extraction have been proposed. In this section, we describe some of the linear feature extraction methods, i.e., the extracted features can be written as linear combinations of the original features. Nonlinear feature extraction techniques are more sophisticated. The readers may also find two recent surveys [55,56] useful in this regard.

Unsupervised Techniques “Unsupervised" here refers to the fact that these feature extraction techniques are based only on the data (pattern matrix), without pattern label information. Principal component analysis (PCA), also known as Karhunen-Loeve Transform or simply KL transform, is arguably the most popular feature extraction method. PCA finds a hyper plane such that, upon projection to the hyper plane, the data variance is best preserved. The optimal hyper plane is spanned by the principal components, which are the leading eigenvectors of the sample covariance matrix. Features extracted by PCA consist of the projection of the data points to different principal components. When the features extracted by PCA are used for linear regression, it is sometimes called “principal component regression". Recently, sparse variants of PCA have also been proposed [20], where

each principal component only has a small number of non-zero coefficients.

Factor analysis (FA) can also be used for feature extraction. FA assumes that the observed high dimensional data points are the results of a linear function (expressed by the factor loading matrix) on a few unobserved random variables, together with uncorrelated zero-mean noise. After estimating the factor loading matrix and the variance of the noise, the factor scores for different patterns can be estimated and serve as a low-dimensional representation of the data.

Supervised Techniques Labels in classification and response variables in regression can be used together with the data to extract more relevant features. Linear discriminate analysis (LDA) finds the projection direction such that the ratio of between-class variance to within-class variance is the largest. When there are more than two classes, multiple discriminate analysis (MDA) finds a sequence of projection directions that maximizes a similar criterion. Features are extracted by projecting the data points to these directions.

Partial least squares (PLS) can be viewed as the regression counterpart of LDA. Instead of extracting features by retaining maximum data variance as in principal component regression, PLS finds projection directions that can best explain the response variable. Canonical correlation analysis (CCA) is a closely related technique that finds projection directions that maximize the correlation between the response variables and the features extracted by projection.

Feature Grouping

In feature grouping, new features are constructed by combining several existing features. Feature grouping can be useful in scenarios where it can be more meaningful to combine features due to the characteristics of the domain. For example, in a text categorization task different words can have similar meanings and combining them into a single word class is more appropriate. Another example is the use of the power spectrum of classification, where each feature corresponds to the energy in a certain frequency range. The preset boundaries of the frequency ranges can be sub-optimal, and the sum of features from adjacent frequency ranges can lead to a more meaningful feature by capturing the energy in a wider frequency range. For gene expression data, genes that are similar may share a common biological pathway and the grouping of predictive genes can be of interest to biologists [21].

The most direct way to perform feature grouping is to cluster the features (instead of the objects) of a data set. Feature clustering is not new; the SAS/STAT procedure “varclus" for variable clustering was written before 1990 [19]. It is performed by applying the hierarchical clustering method on a similarity matrix of different features, which is derived by, say, the Pearson's correlation coefficient. This scheme was probably first proposed in [14], which also suggested summarizing one group of features by a single feature in order to achieve dimensionality reduction. Recently, feature clustering has been applied to boost the performance in text categorization. Techniques based on distribution clustering [15], mutual information and information bottleneck have also been proposed.

Features can also be clustered together with the objects. As mentioned in [18], this idea has been known under different names in the literature, including \bi-clustering" , “co-clustering" , \double-clustering", \coupled clustering" , and \simultaneous clustering" . A bipartite graph can be used to represent the relationship between objects and features, and the partitioning of the graph can be used to cluster the objects and the features simultaneously . Information bottleneck can also be used for this task.

In the context of regression, feature grouping can be achieved indirectly by favoring similar features to have similar coefficients. This can be done by combining ridge regression with LASSO, leading to the elastic net regression algorithm [19]

3. PROPOSED WORK

In the proposed work, first dimensionality of data set is reduced by applying Principal Component Analysis (PCA) and then decision tree is built.

In combination with principal components analysis and Decision tree with their characteristics, firstly, filter the sample data set, then extract the main attributes, and lastly construct a new decision tree by the following algorithm. The detailed is as follows:

Step 1 In starting step we Convert data source into a multi-matrix, identify the main attributes by principal components analysis.

1) Get mean vector of all the feature vectors (xmean).
2) Get x_i-xmean.
3) Get the covariance matrix. (Square and symmetric)

$$cov_{M\times M} = 1/NumofSamples * \left(\sum [x_i - xmean]_{M\times 1} * [x_i - xmean]_{1\times M}{}^T\right)$$

4) Get the Eigen Values and Eigen Vectors.
5) Normalize the Eigen Vectors.
6) Form the Transformation matrix (T) (contains the eigen vectors sorted by putting the eigen vectors that correspond to the max eigen values first).
7) Apply the transformation: y = transpose(T) * [x_i-xmean].
8) Reduction of y :

Calculate the Loss when removing some features of y.

- Sort the eigen values discerningly and remove from the smaller (from the bottom).

$$Loss = \frac{\sum \lambda(to\ be\ removed)}{\sum \lambda(all)}$$

Remove the features according to the required loss (make them zeroes) → get y.

→This way, we reduced the number of features.

Step2 Do data cleaning for data source and generate the training sets of decision tree through converting the continuous data into discrete variables.

Step3 Compute the information (entropy) of training sample sets, the information (entropy) of each attribute, split information, information gain and information gain ratio, of which S stands for the training sample sets and A denotes the attributes.

Step4 Each possible value of root may correspond to a subset. Do step 3 recursively and generate a decision tree for the sample subset until the observed data of each divided subset are the same in the classification attributes.

Step5 Extract the classification rules based on the constructed decision tree and do classification of new data sets.

4. Conclusion

Classification algorithm is used in various applications as we have discussed above but it is not limited, the algorithm is also useful for natural disasters like cloud bursting, earthquake etc. A Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possible correlated variables into a set of values of linearly uncorrelated variables for dimension reduction to enhance the classification rate.

5. REFERENCES

[1] Hughes and L. Tarassenko, "Novel Signal Shape Descriptors Through Wavelet Transforms And Dimensionality Reduction. In Wavelet Applications In Signal And Image Processing", p.p 763–773, 2003.

[2] S. Venkatarajan and W. Braun, "New Quantitative Descriptors Of Amino Acids Based On Multidimensional Scaling Of A Large Number Of Physicalchemical Properties". Journal of Molecular Modeling, 7(12):445–453, 2004.

[3] C. Jenkins and M.J. Mataric, " Deriving Action And Behavior Primitives From Human Motion Data", In International Conference on Intelligent Robots and Systems, volume 3, p.p 2551–2556, 2002.

[4] Wettschereck, D.W. Aha, and T. Mohri, " A Review And Empirical Evaluation Of Feature Weighting Methods For A Class Of Lazy Learning Algorithms" Artif. Intell. Rev., 11(1-5):273-314, 1997

[5] Mohavi and G. John, " Wrappers For Feature Subset Selection. Artificial Intelligence", 97(1-2):273-324, 1997.

[6] Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn, and A.K. Jain, " Dimensionality Reduction Using Genetic Algorithms". IEEE Transactions on Evolutionary Computation, 4(2):164-171, July 2000.

[7] Guyon and A. Elissee, " An Introduction To Variable And Feature Selection. Journal Of Machine Learning Research", 3:1157-1182, March 2003.

[8] Tibshirani, " Regression Shrinkage And Selection Via The Lasso. Journal Of The Royal Statistical Society". Series B (Methodological), 58(1):267-288, 1996.

[9] Donoho, " For Most Large Underdetermined Systems Of Linear Equations, The Minimal L^1-Norm Solution Is Also The Sparsest Solution". Technical report, Department of Statistics, Stanford University, 2004

[10] H. Friedman, "Multivariate Adaptive Regression Splines". Annals of Statistics, 19(1):1-67, March 1991.

[11] J.C. MacKay, " Bayesian Non-Linear Modelling For The Prediction Competition". In ASHRAE Transactions, V.100, Pt.2, pages 1053-1062, Atlanta Georgia, 1994.

[12] Kohavi and G. John, " Wrappers For Feature Subset Selection. Artificial Intelligence", 97(1-2):273-324, 1997.

[13] Guyon, S. Gunn, A. Ben-Hur, and G. Dror, " Result Analysis Of The NIPS 2003 Feature Selection Challenge. In Advances In Neural Information Processing Systems", pages 545-552. MIT Press, 2005.

[14] L. Blum and P. Langley, " Selection Of Relevant Features And Examples In Machine Learning. Artificial Intelligence", 97(1-2):245-271, 1997.

[15] Battiti, " Using Mutual Information For Selecting Features In Supervised Neural Net Learning". IEEE Transactions on Neural Networks, 5(4):537-550, July 1994

[16] Kwak and C.-H. Choi, "Input Feature Selection By Mutual Information Based On Parzen Window". IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(12):1667-1671, December 2002.

[17] Torkkola, " Feature Extraction By Non Parametric

Mutual Information Maximization". Journal of Machine Learning Research, 3:1415-1438, March 2003.

[18] L Yu and H. Liu, " Feature Selection For High-Dimensional Data: A Fast Correlation-Based Filter Solution". In Proc. 20th International Conference on Machine Learning, pages 856-863. AAAI Press, 2003.

[19] M A. Hall, " Correlation-Based Feature Selection For Discrete And Numeric Class Machine Learning". In Proc. 17th International Conference on Machine Learning, pages 359-366. Morgan Kaufmann, 2000.

[20] D Koller and M. Sahami, " Toward Optimal Feature Selection". In Proc. 13th International Conference on Machine Learning, pages 284-292. Morgan Kaufmann, 1996.

[21] K Kira and L. Rendell, "The Feature Selection Problem: Traditional Methods And A New Algorithm". In Proc. of the 10th National Conference on Artificial Intelligence, pages 129-134, Menlo Park, CA, USA, 1992. AAAI Press.

[22] I Kononenko, " Estimating Attributes: Analysis And Extensions Of RELIEF". In Proc. 7th European Conference on Machine Learning, pages 171-182, 1994.

[23] R Caruana and D. Freitag, " Greedy Attribute Selection". In Proc. 11th International Conference on Machine Learning, pages 28-36. Morgan Kaufmann, 1994.

[24] R Kohavi and G. John, " Wrappers For Feature Subset Selection". Artificial Intelligence, 97(1-2):273-324, 1997.