A Novel Precise Poverty Alleviation System Based on High Dimensional Data Dimensionality Reduction and Feature Clustering Model

Zhiwei Zhao* and Taiguang Gao

Heilongjiang University of Science and Technology, Heilongjiang, 150022, China

Corresponding author (Email: zzw_zhiwei@126.com)

Abstract: The life standard decides the poverty and this includes the development of the economical environmental and social outcomes of a country or a core region. The spatial distribution information provides a poverty map. Generally census data doesn’t provide required information but it has necessary sample size. Small Surveys provide the poverty map data for income and expenditures. If the census data and survey data are combined the poverty rates can be predicted for the total population. This relating of spatial and non-spatial data on ground provides a map which is utilized for socio economic development. This mixed data becomes a microarray and it expresses the hundreds and lakhs results are produced simultaneously. To analyze this high dimensional large generated data with Gini coefficients an efficient technique is required to cluster with high reduction in data size. The running time is minimized in proposed PCA-ANN and provides more accurate result for the high dimensionality data.

Key words: PCA-ANN, High dimensional data, poverty map, Gini coefficients.

1. INTRODUCTION

The poverty is the factor which mainly depends upon the shortage of the income. The developed and developing countries still continues the eradication of poverty in all facets. The sustained growth of income is the only object to achieve the goal. Only a few developing countries have succeeded in reduction of poverty level. The main priority is given to policies and processes that are essential for the growth of income and reduction of poverty. It is clearly simple to compare the appropriate factors across different regions by empirical analysis. The two methods in analysis strategy, one is mainly based on econometric and the second is based on set of details and using of various tools related to history and socio economic policies. These two strategies are compliment to each other and the average pattern is identified from the data of country studies. The special results of various factors are widely compared and performance (range of variations) is taken for a time period. The strategy may be of anything but in the design process the irrelevant factors should not present, so that the time periods of processing and analyzing of data will be simpler. The cross country regression is has serious limitations in theoretical and real time analysis. It should not be dismissed and for absolute solution of poverty as income growth. The better data extraction betters the inequality between the theoretical and real.

Figure 1. Poverty Alleviation Level
The important thing is to consider that if the analysis policies, import export services and technology must be analyzed in depth without any hazards. If the barrier is removed for any policy it must leads to the increase income inequality at least a region or an individual. The international trade will develops the highly qualified employees and it leads a good stability in economic which are considered as a small transaction unit STU. It mainly consists of transaction between the three levels as individual, industrial and government policies transaction. Second the poor infrastructure must be avoided which affects economy in terms of power, communications, transport, etc., third the poverty measure is done by conventional measures for the data about trends and status in it and it also looks for the another view of poverty of country. By means of different indicating factors a country or a region is considered as successful in developing the standards of people. The factors that lead to this conclusion based on income. A general record shows that around 48.3 percent of the total families are comes under the poverty line. Since the financial lack is due to natural causes also like floods and cyclones. The increase in population increases the poverty over 20 years. The recent census shows that the growth increases much higher than previous years and consumption ratio also increases. This increase in ratio leads to growth in poverty map data. The gap between the rural and urban poverty has been increasing because the decline ratio for rural is only by 3 percentage and for rural cases in has been 18 percentages over a particular period. The in-depth knowledge and blended information is essential for data analysis. Here in poverty alleviation the percentage of poverty is based on data of high dimensionality. The main difference between the traditional data and high dimensional data is their dimensions. The technology development produces the database as much more bigger and complex. The different types of database are trade data, Gini expression data, and frequency data. Other forms of data are related to web and multimedia and it contains more no of data which contains multi dimensions. Due to dimension disaster the high dimensional data becomes complex. The increase in data dimension the performance of index structure is gradually decreased. The similarity concept is available in case of low dimensional data because it uses Euclidean distance for its measurement.

The high dimensional data different algorithms are available to find out the performance and also there is lot of clustering method which fails in producing better efficiency. To solve the problem the high dimensional data must be changed into low dimensional data and then by using the low dimensional approach, without decreasing the efficiency of the algorithm a new incremented parallel algorithm is developed to overcome the increased failure rate problem.

The process starts with the data fragmentation in which single high dimensional data is split into multiple data. These high data which is obtained has the information contains channel description. While in case of divided data it has different types of data as a group such as photos and text files etc. These types of data refer a single item in different angles and produce different information based on our needs. The data becomes large and abundant if it has all the information. For such high data the process of analysis creates another problem such as storage cost. If the distance between two high dimensional data are same resulting the failure of distance measure and the dimensionality curse increases. So the low dimensional data reduces the workload, cost and produce stable smooth efficient output.

The feature selection and extraction is the step by step process in data dimensionality reduction. To select the best feature a mapping method is useful that transform the high dimensional data into low dimensional data. The scattered information is classifies the new and actual features. So a typical feature extraction technique is required. The dimensionality data reduction is the important issue and principal component analysis combined with artificial neural network provides optimum solution for the needs. The other works in data reduction as linear discriminant reduction provides a less efficient than the PCA. In poverty alleviation the data set which decides the poverty line includes

1. Measure based on income
2. Magnitude of poor families
3. Annual Population growth
4. Education
5. Health and Nutrition
6. Access to basic facilities
7. Poverty incidence in Rural and Urban
8. Poverty incidence by Region
9. Human development index

The purpose of data reduction is all about the simplification of collected information, efficient storage and retrieval, noise reduction and based on positive effects the queries are solved accurately. The main applications of dimensionality reduction include micro array analysis, text mining, and classification in proteins, intrusion deduction, face recognition and handwritten digit recognition and the gene expression micro array analysis is proposed here because the gene contains lakhs of DNA and the process is same as these high dimensional data. The total is reduced by analyzing a single cell and that is considered as low dimensional data. The existing statistical methods have advantages and disadvantages. In the linear dimensionality reduction technique has inability to handle the data in multi-dimensional manner. Second the isometric feature mapping is useful for
image processing pattern recognition solutions. Finally the principal component analysis (PCA) and the multi-dimensional scaling (MDS), locally linear embedding are the current approaches followed in data reduction and analysis methods. In this every methods finds a mapping procedure that represents the special and important features of actual data in low dimension with small space. The space between the actual and new data size is related as x<<X.

In classification of clustering based on supervision and unsupervised the learning process of this comes under supervised and the ultimate process is grouping the unclassified data into meaningful clusters. The similarities and dissimilarities must be identified and the data are manually organized. The main draw back in this manually organized process is the cost and time constraints. In order to solve this constraints the machine learning algorithm is used as follows i) Hierarchical clustering algorithms (HCA) ii) Partition based clustering algorithm (PBCA). Based on the similarity the merging and splitting of clusters performed in HCA. While in PBCA the data are grouped into non overlapping partitions. The first one provides a good visualization particularly for the data exist in hierarchy naturally. Subspace clustering is a traditional clustering technique extension in which it finds the clusters from the existing multiple or possibly overlapping clusters. Based on searching strategy the subspace cluster is classified into two approaches one Bottom up approach and another is the top down approach. Top down algorithms makes the use of full set of dimensions and reveal the set of subspaces iteratively and Bottom up approaches consider each object as a separate cluster and combine them to form clusters which starts from an initial set of subspace. The proposed PCA has more advantage than the above when it combined with ANN. The feature selection and extraction was done by PCA in the first stage of data modeling, while artificial neural networks were used to classify data according to their poverty alleviation in the second stage. The modeling strategies were tested continuously to find the better low dimensional data. Based on the international survey done on “poverty alleviation and its strategies to overcome” the data are collected. Various factors describe the education, life style, basic amenities, and essential services, utilizing factors of government schemes by public, social and cultural norms, self-efficacy and other characteristics.

2. THE PROPOSED METHODOLOGY

Principal component analysis (PCA) is a linear transformation which contains actual collected data as variables and that is usually represented into a new set of normally correlated orthogonal values. The main aim in PCA is dimensionality reduction and reduced processing time without affecting noises. The PCA method changes the multidirectional high dimensional data into low dimensional data. It is also useful in common pattern recognition and detection techniques in image processing. The process includes ANN and it is best for the fault identification and classification. Since we are reducing the high dimensional data the possibility of error is more and that can be identified and classified for analysis using this feed forward back propagation neural networks. It consists of an input portion, a hidden portion and an output portion. This three stage structure provides the input pattern for training process and back propagation for error calculation. The back propagation usually begins with a random set of weighted data.

Applying PCA to Reduce the Dimension of the High Dimensional Data Set includes the following.

S1: create a matrix M using the data set.
S2: Normalize the data set using -value.
S3: Calculate the decomposition value of the data matrix.
S4: Calculate the variance using the diagonal elements
S5: Sort variances in decreasing order.
S6: Choose the principal components from with largest variances.
S7: Form the transformation matrix consisting of those values
S8: Find the reduced projected dataset in a new coordinate axis by applying to . The initial centroid value for the data set with dimensionality The variance is computed for each data in dimension. The column with maximum variance is and it sorts it in any order. In the total k subset clusters each median is initialized in the

![Figure 2. Process in Proposed Method and the Structures](image-url)
cluster centers. In linear dimension reduction technique PCA is the best, the mean-square error is based on the covariance matrix of the variables, it is a second-order method. By linear orthogonal combinations of data the largest variables is obtained from the variance. Let \( n = [n_1, n_2, \ldots, n_n] \) be a data vector with \( n \in M \). Then on basis of vectors \( \{\ldots\} \) an orthonormal property is satisfied. The best coordinates in lower dimensional space are defined by dot-products as \( H_{\text{max}} = \). The requirement is: find a direction of \( H \) that maximizes the variance and minimize the mean square error (mse).

The input has \( x \) features and \( m \) number of instances sampled and is the adjustable relevance threshold.

![Figure 3. Feed Forward ANN Architecture](image)

*Initialize the matrix*

*For*

*The output is greater than the focus featured in all data M*

*Initialize the matrix*

*The dimensionality dimension algorithm provides the data output about the dimension for the input and for a given sample \( n = [n1, n2, \ldots, nn] \) where \( n \) is sample size*

*For*  
  //Initialization  
  Generation of class label encoding  
  Initialization  
  // C is Classification  
  Set the maximum number of iterations  
  Set precision, set counter  
  //Dimensionality reduction  
  While  
  Do  
  //  
  \( r = r + 1 \)  
  Obtain an approximate solution with gradient iteration method \( V(r) \)  
  }  
  //print result }
\[ \text{Import} = (l_1, l_2, \ldots, l_n)^T, \omega = \Sigma l_i x_i y_i^T \]

//projection matrix

Calculate the symmetric positive semi definite matrix \( M \), elements, for use the PBB method to solve the optimization problem Constraint conditions \( 0 \)

\[ T \quad \text{Given} \]

If \( l(1) \), \( l(1) \) replace \( (l(1)) \)

Calculate the projection vector ,

Stop the cycle and jump to the final output statement

Calculate

//long of the step

The following is the final output statements

Import:

The algorithmic steps are depicted in figure 3 which shows the process in PCA-ANN and the organization can be demonstrated from the following figure.

![Diagram showing PCA-ANN organization](image)

**Figure 4. Inner Organization of the Proposed Methodology**

The entire process is done in Matlab environment for a large set of data to identify the poverty line. The data are taken based on national censes and survey in the year 2015 by the 2010-2015 Family Income and Expenditures Survey, NSO and poverty ratio is determined by the another data set such as inadequacy of food intake from the survey and the extracted feature using proposed method is shown in the table 1.

<table>
<thead>
<tr>
<th>Table 1. Comparative Analysis Output after Proposed Method Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of a person who got two square meals a day</td>
</tr>
<tr>
<td>Throughout the year</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Rural</td>
</tr>
<tr>
<td>Urban</td>
</tr>
</tbody>
</table>

From the table 1 it is observed that the data shows the exact ratio of poverty on the basis of amenities like food and medicine. In this the urban people less affected than the rural area. This accuracy is obtained by training the data in ANN in which it extracts the large set of data from the survey and census into a exact small set of data which is easy to observe and take the necessary steps. The household ratio for the states are shown in figure 4 in which the two years are compared and the blue line indicated the increased percentage of poverty in every year and red line indicates the less ratio of poverty in year 2012-11 than 2013-14. The knowledge towards
the medicinal treatment is also one of the data which shows the poverty ratio. The awareness in male and female is listed and the results are shown in table 2.

![Figure 5. Household Percentage for States with the Data Mapping and Projection](image)

<table>
<thead>
<tr>
<th></th>
<th>2013-14</th>
<th>2012-13</th>
<th>2011-12</th>
<th>2010-11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rural</td>
<td>48.5</td>
<td>46.2</td>
<td>40.9</td>
<td>35.8</td>
</tr>
<tr>
<td>Urban</td>
<td>75.8</td>
<td>74.2</td>
<td>69.6</td>
<td>54.7</td>
</tr>
</tbody>
</table>

The different 25 sets of data from the NSO, contains 10 types of data sets for income and two 6 types of data sets includes the food adequacy and medicinal scheme awareness. In the each data set, a sample is selected randomly in half samples for training, the remaining samples for testing, the training process repeated for about 120 times. The figure 5 shows the data clusters and different small dimensional data clusters which we got as output after training process.

![Figure 6. Small Dimensional Data Clusters](image)

For better extraction the process repeated again and again and the separated data is shown in figure6. Figure 7 illustrates the cluster segments of the two features namely medical data and its awareness among the group of population clusters taken for the proposed work. Based on the scatter plot, the confusion matrix from the NN architecture is shown in figure 8. The figure 8 shows the trained data by confusion matrix for the set of 10 factors. From the matrix it is observed the G3, G4, G6, G8 has the better percentage CCR.
3. CONCLUSION

The proposed work in this research paper utilizes a feature clustering algorithm with the ANN architecture to address and predict the alleviation of poverty levels on a global basis. Several input features relevant to poverty alleviation have been given as input to the clustering algorithm to obtain the confusion matrix. Based on the literature survey presented in the initial sections, and a cluster based technique has been incorporated to segregate the given data set into clusters to which data reduction technique has been applied to reduce the dimension of the feature vector. This reduction drastically brings down the computation time as well as complexity of the algorithm and an extensive algorithm has been presented in this paper and results justify the efficiency of the proposed method when compared to the conventional techniques. The convergence rate and efficiency without error are obtained by selecting a suitable recognition system. Based on the requirements primitive patterns required pattern grammar is used. The availability of algorithms and enough pre knowledge is required to recognize the pattern. The statistical approach includes many principles and fuzzy is suitable for most of the models. But the cons in the fuzzy can be overcome by the AFSA. The main advantage in proposed method is the pattern recognition process produces output stage by stage even for complex data set.

REFERENCES


