Using PCA and Random Projections to Compare Preference of Performance

Yong-Gyu Jung, Myung Jae Lim* and Young-Jin Choi

Eulji University, Department of Medical IT Marketing, 212 Yangji-Dong Sujung-Gu Sungnam 461-713 Korea. {ygjung, lk04, yuzin}@eulji.ac.kr
**corresponding Author*

Abstract

Datamining is interested and applied in several areas around us. In other words, it is to discover hidden useful correlations and to predict future by extracting the process to make decisions in actionable information. A variety of data can be converted to interpret in other format. Even a simple conversion can find a big difference in analyzing the results. In this paper, the three principal components for analysis and changes were used Random projections and data conversion techniques. The correct classification is compared by performance of two methods for viewing the sample data set using the Bayesian algorithm.

Keywords: *Principal component analysis (PCA), Random projections, data conversion*

1. Introduction

Data mining has been useful in a heap of common sense, even the bulk data by the creation of new knowledge. It is helpful to support decision including the knowledge discovery, so data mining can see the wrong term. In that perspective, KDD is the term commonly used as knowledge discovery from data which are data processing, data mining, data summarization, visualization, machine learning, pattern recognition and knowledge extraction. Simple data processing and analysis of data is several different angles at different points within the meaning. To convert the principal components analysis and Random projections random time series will present a variety of conversion techniques.

In this paper, the principal components analysis with data preprocessing techniques and arbitrary Random projections with the Bayesian algorithm is used to data classification.

2. Related Research

2.1 Data Transformation

To perform a simple transformation can be represented to a big difference. Principal component analysis (PCA) is consisting of a multi-dimensional feature vector data. It is maintained at a lower level without loss of information, while low-dimensional data processing methods to reduce the dimension.

Step 1. Find direction (axis) of greatest variance.

Step 2. Find perpendicular direction of greatest variance to previous direction and repeat.

The algorithm is performed repeatedly for finding the largest distributed direction of the axis such as the dispersion and vertical axis. This is applied to these algorithms when the data vector represents the covariance matrix saved their relationship. The covariance can be obtained for specific vectors and eigenvalues. At this point, the largest eigenvalue

corresponds to the first ingredient which becomes a unique vector principal component. PCA is the high-covariance matrix to configure unique vector transformation. It is to find the 3D time consuming problem, a lot more running data. Random projections are simple work around for this. The average distance of random projections keeps a good relationship. KD-tree and the high-level data can be used to apply to random projections based on other models for improving stability.

2.2 Automatic Data Cleansing

Practical problems concerning the quality are most important matter in datamining. It can be come from the most common errors in large databases. In order to solve this problem, all the datamining technologies, improved decision tree, robust regression, outlier detection can be used as datamining techniques. Improved decision tree is removing some of the incorrectly classified instances. The decision-making techniques are dependent on the size of the tree. Sub-tree is suitable for a statistical test applied to determine whether the decision will be applied locally. In robust regression, the absolute distance measurement is used unlike the standard square in existing linear regression.

Some inaccurate data is rising in the form of automatic discovery. Expert opinion has no way to talk about the shortest error, whether particular instance or type fits the model. Regression is a statistical visualization, even though professionals shall not be visually clear view. For example, We understand the above value is exact in the picture. However, most problems do not resist the easy visualization. In other words, the regression model, the concept of types is more difficult to understand. Although it is a good result in the decision tree as our bodies, It does not fit most standard data set of instances obtained known to the warnings. Particular cases dealing with a new data set is not really convenient for the new data to the decision tree. It is still desirable.

2.3 Using unlabeled Data

Classification is not direct learning of clustering. Recently researchers have investigated the area between the two algorithms called semi-supervised learning.

Step 1. Unlabeled data using the classifier is trained.

Step 2. Label data do not take this class shall be accuracy.

Step 3. To label all of the data used to learn a new classifier.

Step 4. Repeat until you make whatever changes.

It is collected to unlabeled for starting point and the cluster labels from the data. During iteration, it is found the EM process on an equal or better chance guaranteed great model parameters. Only experience can answer the question to improve the accuracy of their category. Co-training from every perspective is the first different models. It is based on the contents and hyperlinks-based model. Then, the labeling of samples is used for each. For each model, positive values have a confidence level of the sample and the negative value of the label to select a sure bet. And this is a sample of some of unlabeled pool is added. Better than by choosing more than one type unlabeled pool to maintain the ratio of the amount of the sample well. Between the labels which are not the entire procedure repeated until the pool is exhausted.

3. Experiments

It has a key role for information technology to increase exponentially with the handling of biological data analysis. Completion of human genome maps published in 2001 that was impossible without the help of information technology. Among them, the protein structure is constantly conducted. The protein consists of 20 amino acids. These amino acids can appear in any order of the amino acid chain. Proteins make up the various structures of the amino acid sequence. It has revealed to meet more than 90%.

In this paper, the main component of the data preprocessing is analysis PCA and random projections. Through the real perform data conversion category, it has any affect in the comparison. As an experimental tool, WEKA is used which was developed at Waikato university with protein data for lesion area named aecoli data. The experimental data is expected changes in the local terminal of the protein, such as determining the expression values of 7 are numeric attributes and values that represent parts of lesions (cytoplasm, cell membranes, in / out, etc.) class appears as eight kinds of databases. In addition, the experiments are repeated with 10 fold validations. Experiments are performed before the actual principal components analysis of the pre-operation and used any way compared Random projections.

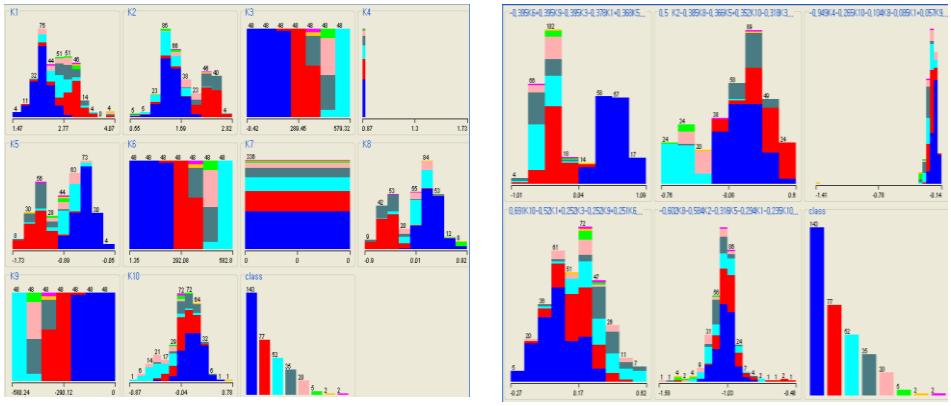

Figure 1. Conversion to Random projections (left) and PCA (right)

Figure 1 represents the attributes applied to an arbitrary random projections, the active ingredient will be filtered visual attributes. If any of the random projections increased the number of attributes compared to the active ingredient can be found to decrease the number of attributes.

4. Conclusion

Currently datamining sector is interested and applied in many areas. Some experimental results obtained from using the information to improve the probability of an event can be the first time. To interpret data on various aspects can be converted to real expectation. Analyzing the results even a simple can be found big difference. There are Principal components analysis and random projections, the property vector text and time series transformation techniques can be useful methods. In this paper, the protein in the lesion area using principal components analysis with data preprocessing techniques and random projections with the Bayesian algorithm are used to analyze the data.

As a result, by analyzing the principal components analysis with a data transformation using any of random projections compared to the performance for the preference of performance. In the future, it will be analyzed and compared performance for conversion technique using automatic data cleaning after preprocessing.

References

[1] G. Shmueli, N. R. Patel and P. Bruce, "Data Mining for Business Intelligence", **(2009)**.
[2] L. H. Witten and E. Frank, "Data.Mining. Practical Machine Learning Tools and Techniques",Third Edition, Morgan Kaufmann **(2011)**.
[3] R. A. Dunne, "A statistical approach to neural networks for pattern recognition", **(2007)**.
[4] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: application to image and text data", Proceedings of the 7the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001), **(2001)**.
[5] Y.-G. Jung, S.-H. Lee and H. J. Sung, "Effective Diagnostic Method Of Breast Cancer Data Using Decision Tree", Journal of IWIT **(2010)**, Vol. 10, No. 5, pp.57-62.
[6] I. C. Kim and Y. G. Jung, "Using Baysian Network to analyze Medical Data", LNAI2734, Springer-Verlag **(2003)**, pp. 317-327.
[7] Y. G. Jung, K. Y. Lee and M. J. Lim, "Discharge Decision for Post-Operative Patients", Proceedings of ICHIT **(2010)**, pp.195-199.

Yong Gyu Jung

1981 BS Seoul National University
1994 Master of Engineering, Yonsei University
2003 Doctor of Science, Kyonggi University
1999-present, professor in the Department of IT marketing, Eulji University
<Interest areas: Clinical Data Mining, Medical Information Systems, EDI Standards (UNEDIFACT, ebXML)>

Myung Jae Lim

1989 BS Chung-Ang University
1992 Master of Engineering, Chungang University
1998 Doctor of Engineering, Chungang University
1992-present, professor in the Department of IT marketing, Eulji University
<Interest areas: u-Healthcare system, Information Retrieval, Human Computer Interaction, SW development Methodology>

Young-Jin Choi

1988 Master of Business Administration, Hankuk University of Foreign Studies
2004 Doctor of Business Administration, Sungkyunkwan University
2006-present professor in the Department of Healthcare Management, Eulji University
<Interest areas: IT Governance, Medical Information Systems>