International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.4 Issue No.3, pp : 146-148 01 March. 2015

A Comparative Study of Various Data Transformation Techniques in Data Mining

Km. Swati, Dr. Sanjay Kumar
Department of Computer Science and Engineering, Jaipur National University, Jaipur, India

Abstract: This research paper presents a technique to select an ideal transformation technique of original and transformed features. The paper reviews about a comparative study of various data transformation techniques used in data mining which includes six types of transformation techniques - Wavelets, Genetic Algorithm and Wrappers, Identity transform, Program synthesis, Data refinement transformation, and Feature Selection technique. The feature selection technique is considered best as it utilizes Wavelets and Genetic Algorithm and Wrappers methods that employ classification accuracy as its fitness function. The selection of transformed features provides new insight on the interactions and behaviors of the features. This method is especially effective with temporal data and provides knowledge about the dynamic nature of the process. The comparative study from the feature selection technique demonstrates an improvement in classification accuracy, reduction in the number of rules, and decrease in computational time.
Keywords: Data transformation ,wavelets, genetic algorithm and wrappers, feature selection technique

Introduction:

Data transformation converts a set of data values from the data format of a source data system into the data format of a destination data system. It is the process of converting data from one format (e.g. a database file or Excel sheet) to another. Because data often resides in different locations and formats across the enterprise, data transformation is necessary to ensure data from one application or database is intelligible to other applications and databases, a critical feature for applications integration. In a typical scenario where information needs to be shared, data is extracted from the source application or data warehouse, transformed into another format, and then loaded into the target location. The quality of knowledge extracted from a data set can be enhanced by its transformation. Discretization and filling missing data are the most common forms of data transformation.

It is an integral part of data mining and knowledge discovery [1] so often used in data mining. For example, data is often normalized to improve the effectiveness of the learning algorithms (clustering, neural networks), but the effects of data transformation on classification accuracy and knowledge discovery has been limited.
Transforming data allows for an increased understanding of the data and discovery of new and interesting relationships between features.

Data transformation removes noise from data and also summarizes data. Data transformation operations, such as normalization and aggregation are additional data preprocessing procedures that would contribute towards the success of mining process. The Data transformation techniques can be utilized with temporal data to improve the quality of knowledge.
Applying these transformations increases classification accuracy of the extracted knowledge, enhances understanding of the behavior of the features which results in more generalizable rules sets (i.e., reduction
of the number of rules), and decreases the computation time.

VARIOUS DATA TRANSFORMATION TECHNIQUES

A) Wavelets

Wavelet transformations were developed to express the frequency domain and the time locality of an input function. The fact that wavelets capture the temporal nature of the data is quite essential.Wavelet transformations consist of a "family" of functions [2].
The wavelet transform is a tool that divides up data, functions, or operators into different frequency components and then studies each component with a resolution matched to its scale [3].
Therefore, the wavelet transform is anticipated to provide economical and informative mathematical representation of many objects of interest [4]. Nowadays many computer software packages contain fast and efficient algorithms to perform wavelet transforms. Due to such easy accessibility wavelets have quickly gained popularity both in theoretical research and in applications. Above all, wavelets have been widely applied in such computer science research areas as image processing, computer vision, network management, and data mining.

B) Genetic Algorithm and Wrappers

A genetic algorithm is a search technique that is based on natural systems. It is used as a classifier directly in computation and also optimize the results. Most applications of genetic algorithm in pattern recognition optimize some parameters in the classification process [5].
Genetic algorithms has been applied to find an optimal set of feature weights that improve classification accuracy. First, a traditional feature extraction method such as Principal Component Analysis (PCA) is applied, and then a classifier such as k-NN (Nearest Neighbor Algorithm) is used to calculate the fitness function for it [6], [7]. Combination of classifiers is another area that Genetic Algorithms have been used to optimize. It is also used in selecting the prototypes in the case-based classification.

The second method of genetic algorithm to optimize the result from the dataset is more effective to compute the accurate values of observations of data by applying data mining techniques.
A wrapper is a method incorporating a search algorithm and a learning classifier to define ideal feature subsets. The wrapper in this model utilizes a genetic algorithm to produce possible feature subsets and a decision tree to evaluate the quality of each subset. The algorithm was selected because it is widely used as well as it generates implicit knowledge in the form of rules. Various parameters of the optimization can affect the computational resources required to carry out the optimization.
Larger values or population size and number of iterations (see runtime arguments) favor optimal solutions, but increase computational time. This method performs aggressive feature selection that optimizes cross-validation performance. Additionally, it is capable of optimizing any performance measure for any classifier type. Unfortunately, methods using genetic algorithms tend to scale poorly with number of features and training set size.

C) Identity Transform

The identity transform is a data transformation that copies the source data into the destination data without change. The identity transformation is considered an essential process in creating a reusable transformation library.

By creating a library of variations of the base identity transformation, a variety of data transformation filters can be easily maintained.

D) Program synthesis

Synthesis is a special form of automatic programming that is most often paired with a technique for formal verification. The goal is to construct automatically a program that provably satisfies a given high-level specification. In contrast to other automatic programming techniques, the specifications are usually non-algorithmic statements of an appropriate logical calculus.

Data refinement

Data refinement is used to convert an abstract data model (in terms of sets for example) into implementable data structures (such as arrays). Operation refinement converts a specification of an operation on a system into an implementable program (e.g., a procedure).

The post condition can be strengthened and/or the pre condition weakened in this process. This reduces any non determinism in the specification, typically to a completely deterministic implementation.

Feature Selection technique

This technique identifies the best features as compared to others. Several transformation schemes are applied to the original data set. Feature selection is also useful as part of the data analysis process, as it shows which features are important for prediction, and how these features are related. Feature transformation is a process through which a new set of features is created. The variants of feature transformation are feature construction and feature extraction. Both are sometimes called feature discovery.
Feature selection is a term commonly used in data mining to describe the tools and techniques available for reducing inputs to a manageable size for processing and analysis. Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction.
The main objective of feature selection is to choose a subset of input variables by eliminating features, which are irrelevant or of no predictive information. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features.
A genetic wrapper feature selection algorithm is utilized to identify the key features from each of the transformed data sets. The selected features are then combined into a single data set. The genetic wrapper selection method is applied to the combined data set. Wrapper methods use a predictive model to score feature subsets.
Each new subset is used to train a model, which is tested on a hold-out set. Counting the number of mistakes made on that hold-out set (the error rate of the model) gives the score for that subset. As wrapper methods train a new model for each subset, they are very computationally intensive, but usually provide the best performing feature set for that particular type of model.

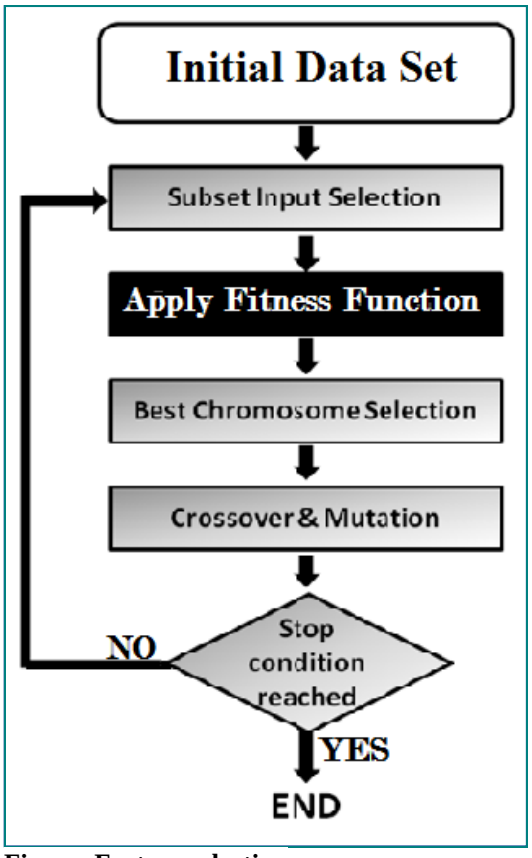

Figure: Feature selection

CONCLUSION

Data mining offer methods and tools for discovery of new knowledge for making decision support system. Classification accuracy of decisions made with the extracted knowledge depends on the properties of the training data set. In most data mining applications raw data is used for rule and feature extraction.

Feature Selection technique identifies the best features as compared to others. The comparative study from the feature selection technique demonstrates an improvement in classification accuracy, reduction in the number of rules, and decrease in computational time. The methods are useful for dimension reduction when the transformed features have a descriptive power that is more easily ordered than the original features.

In this paper a feature selection transformation method is best transformation technique. This transformation, when applied to a training data set, enhances classification accuracy of the decision rules generated from different-different set. The reason for increased classification accuracy with feature selection might due to the fact that the associations among features and decisions are stronger than those built on feature values.

REFERENCES

i. *Howlett, and L.C. Jain (Eds), Knowledge-Based Intelligent Information and Engineering Systems, LNAI 3213, Vol. I, Springer, Heidelberg, Germany, 2004, pp. 148-154.*

ii. *Hubbard, B.B.: The World According to Wavelets: The Story of a Mathematical Technique in the Making. 2nd edn. Peters, A.K. (eds.) Natick, MA (1998)*

iii. *I. Daubechies. Ten Lectures on Wavelets. Capital City Press, Montpelier, Vermont, 1992.*

iv. *F. Abramovich, T. Bailey, and T. Sapatinas.Wavelet analysis and its statistical applications. JRSSD, (48):1–30, 2000.*

v. *Pattern Recognition Letters, (1995).Vol. 16, pp. 801-808*

vi. *Siedlecki, W., Sklansky J., A note on genetic algorithms for large-scale feature selection, Pattern Recognition Letters, Vol. 10, Page 335-347, (1989).*

vii. *Pei, M., Punch, W.F., and Goodman, E.D. "Feature Extraction Using Genetic Algorithms".*