Feature Transformation and Subset Selection

Huan Liu
Dept. of Info. Sys. & Comp. Sci.
National Univ. of Singapore
Singapore 119260
liuh@iscs.nus.edu.sg

Hiroshi Motoda
The Inst. of Scientific and Industrial Research
Osaka University
Mihogaoka, Ibaraki, Osaka 567, Japan
motoda@sanken.osaka-u.ac.jp

Keywords: feature selection, feature transformation, feature extraction, feature construction, machine learning, inductive learning

1 Background

As computer and database technologies constantly advance, human beings rely more and more on computers to accumulate data, process data, and make use of data. Machine learning, knowledge discovery, and data mining are some of Artificial Intelligence (AI) tools that help mankind accomplish those tasks. Researchers and practitioners realize that in order to use these tools effectively, an important part is pre-processing in which data is processed before it is presented to any learning, discovering, or visualizing algorithm. In many discovery applications (for example, marketing data analysis), a key operation is to find subsets of the population that behave enough alike to be worthy of focused analysis [1]. Although many learning methods attempt to either select/extract or construct features, both theoretical analyses and experimental studies indicate that many algorithms scale poorly to domains with large numbers of irrelevant and/or redundant features [4]. All the evidence suggests the need for additional methods for this purpose. Feature transformation and subset selection are some frequently used techniques in data pre-processing.

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. Assuming the original set consists of $A_1, A_2, ..., A_n$ features, these variants can be defined below.

Feature construction is a process that discovers missing information about the relationships between features and augments the space of features by inferring or creating additional features [5, 7, 6]. After feature construction, we may have additional m features $A_{n+1}, A_{n+2}, ..., A_{n+m}$. For example, a new feature A_k ($n < k \leq n+m$) could be constructed by performing a logical operation on A_i and A_j from the original set. Another example is: a two-dimensional problem (say, A_1=width and A_2=length) may be transformed to a one-dimensional problem (B_1=area) after B_1 is discovered.

Feature extraction is a process that extracts a set of new features from the original features through some functional mapping [8]. After feature extraction, we have $B_1, B_2, ..., B_m$ ($m < n$), $B_i = F_i(A_1, A_2, ..., A_n)$, and F_i is a mapping function. For instance, $B_1 = c_1 A_1 + c_2 A_2$ where c_1 and c_2 are coefficients.

Subset selection is different from feature transformation in that no new features will be generated, but only a subset of original features is selected and the feature space is reduced [3, 2]. As to feature transformation, feature construction often expands the feature space, whereas feature extraction usually reduces the feature space.

Feature transformation and subset selection are not two totally independent issues. For example, feature construction and subset selection can be viewed as two sides of the representation problem. We can consider features as a representation language. In some cases where this language contains more features than necessary, subset selection helps simplify the language; in other cases where this language is not sufficient to describe the problem, feature construction helps enrich the language. It is common that some constructed features are not useful at all. Subset selection can then remove these useless features. It is also common to see the combined use of feature extraction and subset selection. Some articles in this issue skillfully combine various feature transformation and subset selection methods. How feature transformation and subset selection are targeted depends on the purpose, i.e. whether it is for concept description or for classification. The former aims at preserving the topological structure of the data whereas the latter aims at enhancing the predictive power.

2 Objective and Scope

There is broad interest in feature transformation and subset selection among practitioners from Statistics, Pattern Recognition, Data Mining, and Knowledge Discovery to Machine Learning since data pre-processing is an essential step in the knowledge discovery process for real-world applications.

The objective of this special issue is to report on the recent studies in feature transformation and subset selection. The main goal is to increase the awareness of the AI community to the research of feature transformation and subset selection, currently conducted in isolation. Through this issue, we hope to produce a contemporary overview of modern solutions, to create synergy among these seemingly different branches but with a similar goal - facilitating data processing and knowledge discovery, and to point to future research directions.

This special issue covers various aspects of feature transformation and subset selection:

- General description of problems with feature transformation and subset selection;

- New methods of feature extraction, construction, and selection;
- Applications of feature transformation and subset selection to real-world problems;
- Combinations of different methods like machine learning, statistics, neural networks, and genetic algorithms in feature transformation and selection.

3 Current Status of Feature Transformation and Selection

This issue reflects the state-of-art in feature transformation and subset selection in many aspects. The articles included are quite representative and completed with references. By reading these articles, readers can get a grasp of what it is about feature transformation and subset selection. Start from there and following the references included, you will find an extensive and active field about feature transformation and subset selection.

3.1 Combination of Feature Transformation and Subset Selection

Instead of considering each individually, researchers now explore the advantages of both feature transformation and subset selection. For example, Bloedorn and Michalski investigated feature selection, quantization and construction as operators in their data-driven constructive induction. The order of applying these operators can be either specified by a user or predefined. It is still premature to draw general conclusions about the best ordering. Lavrač et. al. first expanded the initial features and then pruned the irrelevant ones, both operations are performed systematically in an inductive logic programming setting. Vafaie and De Jong applied genetic algorithms to both feature selection and construction. The two operators are used in turn to search for adequate but necessary features through appropriate selection and construction. Zupan et. al. suggested to use function decomposition to identify appropriate and redundant subsets of existing features, eliminated these redundant features and discover a hierarchy of new features that can be added into the existing feature set.

3.2 Real-world applications

Zupan et. al. applied feature transformation to allocating housing loans. Their feature transformation program HINT discovered feature structures meaningful to a domain expert. Vafaie and De Jong described in detail their experiments of feature transformation with the face data from the FERET database. The number of features and error rates are significantly reduced. Bloedorn and Michalski showed positive effects of feature selection, discretization and construction to two real-world problems: text categorization and natural scene interpretation. Pudil and Novovičová applied their method to texture discrimination problem and speech recognition problem and showed that the method works very effectively even without knowledge of underlying probability structure.

3.3 Using knowledge

Although the majority of the papers have a data-driven flavor, several systems do bear in mind that (a) different problems render various approaches necessary, (b) domain knowledge helps effective search of appropriate features. This type of consideration can be clearly seen in Pudil and Novovičová's work. In addition, Bloedorn and Michalski allow a user to define the order how the three operators should be applied. Yang and Honavar pointed out that their multi-criteria fitness function can be improved by using domain knowledge.

3.4 Complimentary coverage

Six selected papers touch upon various technologies for feature transformation and subset selection, which includes pattern recognition, statistics, machine learning, neural networks, and genetic algorithms. A good example of complimentary coverage is the two papers using genetic algorithm. In Vafaie and De Jong's work, they explain in detail how feature construction and selection are done in a genetic algorithm; in order to speed up the calculation of the fitness function, they used C4.5's accuracy as the measure, which renders the selected and constructed features improper for non-linear classifiers to use, e.g. neural network. Yang and Honavar suggest, in a complimentary way, to use a relatively fast inter-pattern distance-based neural network learning algorithm in the fitness function. They also discussed the issue of multi-criteria optimization which often occurs in practice where both accuracy and cost of classification should be considered. Feature transformation includes principal component analysis, linear discrimination, Fourier transformation, etc. that have been well studied in traditional engineering. Zupan et. al., however, approaches to feature transformation from another angle, via function decomposition.

4 What is lacking and future work

Several papers have made initial attempts to combine different learning algorithms or statistics work and to combine feature construction and selection. More work is needed in unifying this currently diversified field.

4.1 Data categorization

- We need to go beyond the UC Irvine datasets for applications in knowledge discovery from databases. The data we usually use are boolean, nominal, numeric, or relational. Another type of data found in applications is temporal (time sequence prediction, e.g.). In general, numeric data has dimensions (e.g., time, mass, pressure) and scales (e.g., nominal, ratio, interval). Knowing this helps deepen the understanding of the underlying phenomena and exclude irrational combination when to construct new features.

- We need to investigate relations between different data types so that existing technologies for one type of data can be applied to another. Doing so, we equivalently make more methods available in handling various applications.

- We need to study the links between various technologies and data types as experimented in Pudil and Novovičová's work. This is essential since the choices of methods will only become more and more as technologies advance.

4.2 Perspectives from other disciplines

Feature transformation and selection is not an issue only particular to machine learning. Other fields do have a similar problem which may not be phrased as such. For example, pattern recognition in statistics has a long history. The emphases there were more on the numeric data. Switching circuit design in electrical engineering implicitly addresses the feature transformation. One of the motivations of organizing this issue is to create the synergy between different approaches and cross-paradigm connections. This effect can be clearly seen in this issue. Several papers have made initial attempts to combine different learning algorithms or statistics work and to combine feature construction and selection. Can we do more? By examining the approaches from other fields to the same or similar problem and characterizing individual approaches, we can incorporate those into the current repertoire and make feature transformation and selection techniques more general to use and as such we advocate application of machine learning. What is also important is to avoid re-inventing wheels and to take advantages of the available technologies. The latter is particularly important when we solve real world problems.

References

[1] R.J. Brachman and T. Anand. The process of knowledge discovery in databases: A human-centered approach. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 37–57. AAAI Press / The MIT Press, 1996.

[2] M. Dash and H. Liu. Feature selection methods for classifications. *Intelligent Data Analysis: An International Journal*, 1(3):forthcoming, 1997.

[3] P. Langley. Selection of relevant features in machine learning. In *Proceedings of the AAAI Fall Symposium on Relevance*. AAAI Press, 1994.

[4] P. Langley. *Elements of Machine Learning*. Morgan Kaufmann, 1996.

[5] C Matheus. The need for constructive induction. In L.A. Birnbaum and Collins G.C., editors, *Machine Learning - Proceedings of the Eighth International Workshop*, pages 173–177, June 1991.

[6] C.J. Thornton. *Techniques of Computational Learning: an introduction*. Chapman and Hall, 1992.

[7] J. Wnek and R.S. Michalski. Hypothesis-driven constructive induction in aq17-hci: A method and experiments. *Machine Learning*, 14, 1994.

[8] N. Wyse, R. Dubes, and A.K. Jain. A critical evaluation of intrinsic dimensionality algorithms. In E.S. Gelsema and Kanal L.N., editors, *Pattern Recognition in Practice*, pages 415–425. Morgan Kaufmann Publishers, Inc., 1980.

Biographical Sketch

Huan LIU is a senior lecturer in Department of Information Systems and Computer Science at National University of Singapore. He was with Telecom Australia Research Labs from November 1989 to January 1994. He obtained his PhD (Dec. 1989) and MSc (Dec. 1985) in Computer Science at University of Southern California, his B.Eng (July 1983) in Electric Engineering and Computer Science at Shanghai Jiao Tong University. He is senior member of IEEE and Computer Society, member of ACM and AAAI. He has published extensively in Intelligent Systems Design, Telecommunications Systems Design, Neural Networks and Symbolic Computation, Knowledge Acquisition and Data Preprocessing. He is Co-Guest Editor for IEEE Expert Special Issue (forthcoming) on "Feature Transformation and Subset Selection", was co-editor of two books, co-chair of 1993 Australia Joint Conference on AI, and has served on program committees for several international conferences.

Hiroshi Motoda is a professor in the division of Intelligent Systems Science at the Institute of Scientific and Industrial Research of Osaka University since 1996. Before joining the university, he had been with Hitachi since 1967, participated as a senior researcher in research on core management, control and design of nuclear power reactors and expert systems for plant diagnosis at the Energy Research Laboratory, and later reached the position of a senior chief research scientist at the Advanced Research Laboratory where he headed an AI group and conducted research on machine learning, knowledge acquisition, diagrammatic reasoning and information filtering. His current research interests includes, in addition to these, scientific knowledge discovery and data mining.

He received his Bs (1965), Ms (1967) and PhD (1972) degrees in nuclear engineering from university of Tokyo. He was on the board of trustee of the Japan Society of Software Science and Technology and the Japanese Society for Artificial Intelligence (JSAI), the chair of SIG-KBS (Knowledge Based Systems) of JSAI, and on the editorial board of AI and Knowledge Acquisition and IEEE Expert. He is now on the board of trustee of the Japanese Cognitive Science Society (JCSS) and on the editorial board of Artificial Intelligence in Engineering, International Journal of Human-Computer Studies. He is a member of AAAI, IEEE Computer Society, JSAI, JSSST, IPSJ and JCSS. He received the best paper awards twice from Atomic Energy Society of Japan, and twice from Japanese Society for Artificial Intelligence.