Encyclopedia of Data Warehousing and Mining

Second Edition

John Wang
Montclair State University, USA

Volume II
Data Pro–I

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content: Kristin Klinger
Director of Production: Jennifer Neidig
Managing Editor: Jamie Snavely
Assistant Managing Editor: Carole Coulson
Typesetter: Amanda Appicello, Jeff Ash, Mike Brehem, Carole Coulson, Elizabeth Duke, Jen Henderson, Chris Hrobak, Jennifer Neidig, Jamie Snavely, Sean Woznicki
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of data warehousing and mining / John Wang, editor. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
Summary: "This set offers thorough examination of the issues of importance in the rapidly changing field of data warehousing and mining"--Provided by publisher.
ISBN 978-1-60566-010-3 (hardcover) -- ISBN 978-1-60566-011-0 (ebook)
1. Data mining. 2. Data warehousing. I. Wang, John,
QA76.9.D37E52 2008
005.74--dc22
2008030801

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

On Interacting Features in Subset Selection

Zheng Zhao
Arizona State University, USA

Huan Liu
Arizona State University, USA

INTRODUCTION

The high dimensionality of data poses a challenge to learning tasks such as classification. In the presence of many irrelevant features, classification algorithms tend to overfit training data (Guyon & Elisseeff, 2003). Many features can be removed without performance deterioration, and feature selection is one effective means to remove irrelevant features (Liu & Yu, 2005). Feature selection, also known as variable selection, feature reduction, attribute selection or variable subset selection, is the technique of selecting a subset of relevant features for building robust learning models. Usually a feature is relevant due to two reasons: (1) it is strongly correlated with the target concept; or (2) it forms a feature subset with other features and the subset is strongly correlated with the target concept. Optimal feature selection requires an exponentially large search space ($O(2^n)$, where n is the number of features) (Almual-lim & Dietterich, 1994). Researchers often resort to various approximations to determine relevant features, and in many existing feature selection algorithms, feature relevance is determined by correlation between individual features and the class (Hall, 2000; Yu & Liu, 2003). However, a single feature can be considered irrelevant based on its correlation with the class; but when combined with other features, it can become very relevant. Unintentional removal of these features can result in the loss of useful information and thus may cause poor classification performance, which is studied as attribute interaction in (Jakulin & Bratko, 2003). Therefore, it is desirable to consider the effect of feature interaction in feature selection.

BACKGROUND

The goal of feature selection is to remove irrelevant features and retain relevant ones. We first give the definition of feature relevance as in (John et al., 1994).

Definition 1 (Feature Relevance):

Let F be the full set of features, F_i be a feature and $S_i = F - \{F_i\}$. Let $P(C|S)$ denote the conditional probability of class C given a feature sets. A feature F_i is relevant *iff*

$$\exists S_i' \in S_i, \text{ such that } P(C \mid F_i, S_i') \neq P(C \mid S_i') \qquad (1)$$

Definition 1 suggests that a feature can be relevant, if its removal from a feature set reduces the prediction power of the feature set. A feature, whose removal does not reduce the prediction power of any feature set, is an irrelevant feature and can be removed from the whole feature set without any side-effect. From Definition 1, it can be shown that a feature can be relevant due to two reasons: (1) it is strongly correlated with the target concept; or (2) it forms a feature subset with other features and the subset is strongly correlated with the target concept. If a feature is relevant because of the second reason, there exists feature interaction. Feature interaction is characterized by its irreducibility (Jakulin & Bratko, 2004). We give the definition of *k*th-order below.

Definition 2 (*k*th order Feature Interaction):

Let F be a feature subset with k features $F_1, F_2, \ldots, F_k$. Let $\Im$ denote a metric that measures the relevance of a feature or a feature subset with the class label. Features $F_1, F_2, \ldots F_k$ are said to interact with each other *iff*: for an arbitrary partition $S = \{S_1, S_2, S_3, \ldots, S_l\}$ of F, where $2 \leq l \leq k$ and $S_i \neq \phi$, we have $\forall i \in [1, l]$, $\Im(F) > \Im(S_i)$.

It is clear that identifying either relevant features or *k*th-order feature interaction requires exponential time. Therefore Definitions 1 and 2 cannot be directly applied to identify relevant or interacting features when the dimensionality of a data set is huge. Many efficient feature selection algorithms identify relevant features based on the evaluation of the correlation between the

class and a feature (or a current, selected feature subset). Representative measures used for evaluating feature relevance includes (Liu & Motoda, 1998): distance measures (Kononenko, 1994; Robnik-Sikonja & Kononenko, 2003), information measures (Fleuret, 2004), and consistency measures (Almuallim & Dietterich, 1994), to name a few. Using these measures, feature selection algorithms usually start with an empty set and successively add "good" features to the selected feature subset, the so-called sequential forward selection (SFS) framework. Under this framework, features are deemed relevant mainly based on their individually high correlations with the class, and relevant interacting features of high order may be removed (Hall, 2000; Bell & Wang, 2000), because the irreducible nature of feature interaction cannot be attained by SFS. This motivates the necessity of handling feature interaction in feature selection process.

MAIN FOCUS

Finding high-order feature interaction using Definitions 1 and 2 entails exhaustive search of all feature subsets. Existing approaches often determine feature relevance using the correlation between individual features and the class, thus cannot effectively detect interacting features. Ignoring feature interaction and/or unintentional removal of interacting features might result in the loss of useful information and thus may cause poor classification performance. This problem arouses the research attention to the study of interacting features. There are mainly two directions for handling feature interaction in the process of feature selection: using *information theory* or through *margin maximization*.

Detecting Feature Interaction via Information Theory

Information theory can be used to detect feature interaction. The basic idea is that we can detect feature interaction by measuring the information loss of removing a certain feature. The measure of information loss can be achieved by calculating *interaction information* (McGill, 1954) or McGill's multiple mutual information (Han, 1980). Given three variables, *A, B* and *C*, the interaction information of them is defined as:

$$I(A;B;C) = H(AB) + H(BC) + H(AC) - H(A) - H(B) - H(C) - H(ABC)$$
$$= I(A,B;C) - I(A;C) - I(B;C) \quad (2)$$

Here $H(\cdot)$ denotes the entropy of a feature or a feature set. $I(X; Y)$ is the mutual information between X and Y, where X can be a feature set, such as $\{X_1, X_2\}$. Interaction information among features can be understood as the amount of information that is common to all the attributes, but not present in any subset. Like mutual information, interaction information is symmetric, meaning that I(A;B;C) = I(A;C;B) = I(C;B;A) = …. However, interaction information can be negative.

If we set one of the features in the interaction information to be the class, then the interaction information can be used to detect the *2*-way feature interaction as defined in Definition 2. $\Im$, the metric that measures the relevance of a feature or a feature subset with the class label, is defined as the mutual information between a feature or a feature set and the class. Positive interaction information indicates the existence of interaction between features.

Using the interaction information defined in Formula (2), we can only detect *2*-way feature interaction. To detect high order feature interaction, we need to generalize the concept to interactions involving an arbitrary number of attributes. In (Jakulin, 2005) the *k-way interaction information* is defined as:

$$I(S) = -\sum_{T \subseteq S} (-1)^{|S \setminus T|} H(T) \quad (3)$$

Where S is a feature set with k features in it, T is a subset of S and "\" is the set difference operator. $|\cdot|$ measures the cardinality of the input feature set, and $H(T)$ is the entropy for the feature subset T and is defined as:

$$H(T) = -\sum_{v \in T} P(v) \log_2 P(v) \quad (4)$$

According to Definition 3, the bigger the $I(S)$ is, the stronger the interaction between the features in S is. The k-way multiple mutual information defined in (Jakulin, 2005) is closely related to the lattice-theoretic derivation of multiple mutual information (Han, 1980), $\Delta h(S) = -I(S)$, and to the set-theoretic derivation of multiple mutual information (Yeung, 1991) and co-information (Bell, 2003) as $I'(S) = (-1)^{|S|} \times I(S)$.

Handling Feature Interaction via Margin Maximization

Margin plays an important role in current research of machine learning. It measures the confidence of a classifier with respect to its predictions. The margin of a classifier can be defined by two ways: sample-margin and hypothesis-margin. Sample-margin, as defined in Support Vector Machine (SVM) (Vapnik 1995), measures the distance between an instance and the decision boundary obtained from the classifier. While, the hypothesis-margin measures the "distance" between two alternative hypotheses (or predictions) which may be derived by a classifier on a certain instance. For classifiers, such as SVM (sample margin) and Adaboost (Freund & Schapire 1997) (hypothesis margin), a large margin ensures high accuracy as well as good generalization capability. Recall the fact that the removal of interacting features results in the loss of useful information and thus cause poor classification performance. It is quite straightforward that we can (explicitly) handle the feature interaction problem by selecting a subset of features that ensure the acquisition of large margin for classifiers. Apparently wrapper model (Koha & John 1997) can be utilized to achieve the margin maximization in feature selection. However, because of the high computation expense and its nature of inheriting bias from the classifier used in the wrapper model, a filter model is usually more preferable for feature selection algorithm design (Blum & Langley 1997).

Mostly, hypothesis margin is used to design selection algorithms in this category. Comparing with sample margin, the advantages of hypothesis margin are: first hypothesis margin is easier for computation, and second, hypothesis margin lower bounds the sample margin. In (Bachrach-et al. 2004), the hypothesis-margin of an instance x with respect to a set of points P is defined as:

$$\theta_P(x) = \frac{1}{2}(\|x - nearmiss(x)\| - \|x - nearhit(x)\|) \quad (5)$$

Here *nearhit(x)* and *nearmiss(x)* denote the nearest point to x in P with the same and different label respectively, and $\|\cdot\|$ is a distance measurement. On an instance x, the hypothesis margin measures the robustness of the prediction of the Nearest Neighbor classifier on the instance. It is clear-cut that a negative or small positive margin means an error or an unstable prediction, and a large positive margin means a correct and stable prediction. Therefore, as the sum of the hypothesis margins on each points, a large positive hypothesis margin on the whole dataset ensures the small error rate as well as the robustness of prediction. Based on this definition, a family of algorithms for feature selection using filter model can be developed (Robnik-Sikonja & Kononenko, 2003; Bachrach-et al. 2004; Navot-et al. 2005). RELIEF (Kira & Rendell, 1992; Robnik-Sikonja & Kononenko, 2003), one of the most successful feature selection algorithms, can be shown to be in this category. Figure 1 shows the RELIEF algorithm.

In Figure 1, w is an N dimension vector whose i-th element corresponding to the i-th feature. T is the total sampling times and N is the number of features. RELIEF tries to update the weight of features according to their contribution of the hypothesis margin on points of S. Therefore measured on the feature subset selected

Figure 1. the RELIEF algorithm

1. initialize the weight vector to zero: $w = 0$
2. for $t = 1 \ldots T$,
   (a) pick randomly an instance x from S
   (b) for $i = 1 \ldots N$,
   $$w_i = w_i + (x_i - nearmiss(x)_i)^2 - (x_i - nearhit(x)_i)^2$$
3. the chosen feature set is $\{i|w_i > \tau\}$ where τ is a threshold

by RELIEF, the hypothesis margin on the dataset S is big. Obviously, in RELIEF, the hypothesis margin maximization is achieved by online optimization (Krumke, 2002).

FUTURE TRENDS

Though feature interaction can be detected by interaction information, detecting high-order feature interactions is still a daunting task, especially when the dimensionality of the data is huge. In the framework of feature selection via margin maximization, feature interaction is handled efficiently, but its detection is implicit. Hence, it is highly desirable to investigate and develop efficient algorithms and new measurements that can effectively detect high-order feature interaction. One possible way is to combine the two frameworks into one, in which a margin-based feature selection algorithm, such as RELIEF, is first used to reduce the original input space to a tractable size, and then using the interaction information to detect feature interaction in the reduced feature space. In real world applications, the detection of interacting features goes beyond accurate classification. For example, in discriminant gene identification, finding interactions between genes involved in common biological functions is a way to get a broader view of how they work cooperatively in a cell. Therefore, it is promising to use feature interaction detection approaches to assist people to acquire better an understanding about the real-world problems in which interacting features exist.

CONCLUSION

Interacting features usually carry important information that is relevant to learning tasks. Unintentional removal of these features can result in the loss of useful information, and eventuate poor classification performance. Detecting feature interaction especially for data with huge dimensionality is computationally expensive. The challenge is to design efficient algorithms to handle feature interaction. In this short article, we present a brief review of the current status and categorize the existing approaches into two groups: *feature interaction detecting via information theory* and *handling feature interaction via margin maximization*. As the demand for finding interacting features intensifies, we anticipate the burgeoning efforts in search of effective and efficient approaches to answer pressing issues arising from many new data-intensive applications.

REFERENCES

Almuallim, H. & Dietterich, T.G. (1994). Learning Boolean Concepts in the Presence of Many Irrelevant Features. Artificial Intelligence, Elsevier, Amsterdam, 69, 279-305.

Bachrach, R.G.; Navot, A. & Tishby, N. (2004). Margin based feature selection - theory and algorithms. Proceeding of International Conference on Machine Learning (ICML), ACM Press.

Bell, A. J. (2003). The co-information lattice. Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA 2003, edited by S. Amari, A. Cichocki, S. Makino, and N. Murata.

Bell, D.A. & Wang, H. (2000). A formalism for relevance and its application in feature subset selection. Machine Learning, 41, 175-195.

Blum, A.L. & Langley, P. (1997). Selection of Relevant Features and Examples in Machine Learning. Artificial Intelligence, 97, 245-271.

Crammer, K.; Bachrach, R.G.; Navot, A. & Tishby, N. (2002). Margin analysis of the LVQ algorithm. Annual Conference on Neural Information Processing Systems (NIPS).

Fleuret, F.(2004). Fast Binary Feature Selection with Conditional Mutual Information. Journal of Machine Learning Research, MIT Press, 5, 1531-1555.

Freund, Y., and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Computer Systems and Science. 55(1):119–139.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157-1182.

Hall, M.A. (2000). Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning.

Proceeding of International Conference on Machine Learning (ICML), 359-366.

Han, T.S. (1980). Multiple mutual informations and multiple interactions in frequency data. Information and Control, 46(1):26–45.

Jakulin, A. (2005). Machine Learning Based on Attribute Interactions, University of Ljubljana, PhD Dissertation.

Jakulin, A. & Bratko, I. (2003). Analyzing Attribute Dependencies. Proc. of Principles of Knowledge Discovery in Data (PKDD), Springer-Verlag, 229-240

Jakulin, A. & Bratko, I. (2004). Testing the significance of attribute interactions. Proceeding of International Conference on Machine Learning (ICML), ACM Press.

John, G.H.; Kohavi, R. & Pfleger, K. (1994). Irrelevant Feature and the Subset Selection Problem. Proceedings of the Eleventh International Conference on Machine Learning, 121-129.

Kira, K., & Rendell, L. (1992). A practical approach to feature selection. Proceedings of the Ninth International Conference on Machine Learning (ICML), 249–256.

Kohavi, R. & John, G.H. (1997). Wrappers for Feature Subset Selection Artificial Intelligence, 97, 273-324.

Kononenko, I. Bergadano, F. & De Raedt, L. (1994). Estimating Attributes: Analysis and Extension of RELIEF. Proceedings of the European Conference on Machine Learning, Springer-Verlag, 171-182.

Krumke, S.O. (2002). Online Optimization: Competitive Analysis and Beyond. Konrad-Zuse-Zentrum für Informationstechnik Berlin, Technical Report.

Liu, H. & Motoda, H. (1998). Feature Selection for Knowledge Discovery and Data Mining. Boston: Kluwer Academic Publishers.

Liu, H. & Yu, L. (2005). Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering, 17, 491-502.

McGill, W. J. (1954). Multivariate information transmission. Psychometrika, 19(2):97–116.

Navot, A.; Shpigelman, L.; Tishby, N. & Vaadia, E. (2005). Nearest Neighbor Based Feature Selection for Regression and its Application to Neural Activity. Advances in Neural Information Processing Systems (NIPS).

Robnik-Sikonja, M. & Kononenko, I. (2003). Theoretical and empirical analysis of Relief and ReliefF. Machine Learning, 53, 23-69.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc.

Yeung, R. W. (1991). A new outlook on Shannon's information measures. IEEE Trans. On Information Theory, 37:466–474.

Yu, L. & Liu, H. (2003). Feature selection for high-dimensional data: a fast correlation-based filter solution. Proceedings of the twentieth International Conference on Machine Learning (ICML), 856-863.

Zhao, Z. & Liu, H. (2007). Searching for Interacting Features. Proceeding of International Joint Conference on Artificial Intelligence (IJCAI), 1156-1161

KEY TERMS

Feature Selection: An important data preprocessing technique for data mining, helps reduce the number of features, remove irrelevant, redundant, or noisy data, and bring the immediate effects for applications: speeding up a data mining algorithm, improving mining performance such as predictive accuracy and result comprehensibility.

Filter Model: The model relies on general characteristics of the data to evaluate the quality of select features without involving any mining algorithm.

Information Entropy: Information entropy, which is alsocalled Shannon's entropy, is the information-theoretic formulation of entropy. It measures how much information there is in a signal or event. An intuitive understanding of information entropy relates to the amount of uncertainty about an event associated with a given probability distribution. In thermodynamics, entropy is a measurement of the disorder of an isolated system.

Mutual Information: A quantity measures the mutual dependence of variables. It is nonnegative and symmetric.

Margin Maximization: It is a process to find the optimal decision boundary for a classifier, such that the margin (either the hypothesis margin or the sample margin) is maximized.

Support Vector Machines (SVMs): A type of supervised learning methods learns classification boundary through finding optimal hyper-planes between classes.

Wrapper Model: The model requires one predetermined mining algorithm and uses its performance as the evaluation criterion of the quality of selected features. It searches for features better suited to the mining algorithm aiming to improve mining performance