Neurocomputing 92 (2012) 61–68

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Easy-to-explain feature synthesis approach for recommending entertainment video

Tsung-Ju Lee [a], Shian-Shyong Tseng [b,*]

[a] Department of Computer Sciences, National Chiao Tung University, Taiwan
[b] Department of Applied Informatics and Multimedia, Asia University, 500, Lioufeng Rd., Wufeng, Taichung 41354, Taiwan

ARTICLE INFO

Available online 3 March 2012

Keywords:
Dimension reduction
Clustering
Recommendation
Feature synthesis
Unsupervised feature selection

ABSTRACT

The use of dimension reduction techniques has attracted considerable attention owing to information explosion. Without considering the underlying phenomena of interest, traditional dimension reduction approaches aim to search a feature set for optimizing performance. In recommending entertainment videos, beyond the successful recommendations, marketing strategy can be benefited from interpreting precise social context information accurately. Therefore, how to find an easy-to-explain feature set to achieve optimal prediction performance becomes an important issue. In this paper, we propose a three-phase feature synthesis approach to search heuristically optimal feature set within exponential easy-to-explain features. The first phase performs feature selection by screening low-informative features, the second phase shrinks the high-dependent feature subset, and the third phase enhances the dominated features. An implemented social recommendation system and the 11 months purchasing data from the largest commercial entertainment video Web shop in Taiwan are adopted to evaluate the effectiveness and efficiency of the proposed feature synthesis method in the experiments. The experimental results show that our approach can obtain the interpretable clustering results as well as improve the recommendation.

1. Introduction

The problem of dimension reduction has attracted considerable research effort in machine learning and data mining fields. One important goal of dimension reduction techniques is to improve the precision of automatic learning systems as well as to overcome the curse of dimensionality [1]. However, implicit reasons of making decisions dramatically reduce the effects of learning systems. In network security, failing to identify the behaviors of predicted threats in the understandable way of domain experts raises the difficulties of repairing the vulnerabilities. In e-commerce, intelligent social recommendation systems can benefit marketing strategy, such as slotting strategy and shooting plan, if precisely social context can be extracted during the learning phase. The advanced developments of nonlinear learning algorithms decrease these profits while gaining higher precisions.

Traditional dimension reduction techniques which aim to search much less features retaining the key characteristics of data structure can be categorized into feature extraction and feature selection. Feature extraction (also named feature synthesis) algorithms search the best feature subset in much larger feature space generated from original features via particular operations [2–4,6,9,10,11,13,20,22,23]. These linear or nonlinear transformations may produce the implicit features which for a human is difficult to comprehend and hence affect the benefits of intelligent system, even the linear system.

On the other hand, feature selection algorithms search the optimal feature subset of human-understandable features. Due to the computational complexity, various heuristic search algorithms are proposed to earn computational benefit as well as sub-optimal solutions, including *Oscillating search* [16], *random subspace* [8], *local-learning-based algorithms* [18], *evolutionary algorithms* [5], *memetic algorithms* [24], *multiple criterion* [21] and *tabu search* [19]. With the exception of the exhaustive search, optimal solutions are exploited by Branch & Bound algorithms [12,17]. However, these optimal methods can be used only with monotonic criteria. Although these feature selection algorithms maintain the human-understandable features, it usually requires nonlinear learning systems to achieve higher precision due to the complexity of modern data.

One of the easy-to-explain knowledge form is disjunction of conjunctive form. Since the optimization space of disjunction of conjunctive features is much larger than that of original features, the following linear learning system can hence gain better precision while keeping predicted results explainable. In this paper, we propose a three-phase feature synthesis algorithm to search heuristically optimal features within easy-to-explain features generated via logic

* Tel.: +886 4 23323456x1029, +886 920690457; fax: +886 4 23316699.
E-mail address: sstseng@asia.edu.tw (S.-S. Tseng).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.09.034

operators. The first phase performs feature selection by screening low-informative features; the goal of the second phase is to shrink the high-dependent feature subset, and the third phase enhances the dominated features.

To evaluate our feature synthesis algorithm, the entertainment video recommendation is selected as an application. In our experiments, 11 months purchasing data containing 28,249 transactions with 10,906 users are from the largest entertainment video Web shop in Taiwan. The experimental result shows our approach can obtain interpretable clustering results with representative user characteristics and the improvement of our implemented social recommendation system.

The remainder of this paper is organized as follows: In Section 2, we introduce the necessary data mining preliminaries and notations used throughout this paper. Easy-to-explain feature synthesis problem is defined and discussed in Section 3. In Section 4, we proposed our feature synthesis algorithm and the corresponding theoretical benefits. Section 5 presents experiments and performance results in entrainment video recommendation. Our conclusion is given in Section 6.

2. Preliminaries and notations

2.1. Itemsets and notations

In online entertainment video business, folksonomy-based tags are usually used to describe video contents and hence applied to be features of users' preferences. Unlike traditional data mining, each transaction contains only one purchased video and the corresponding annotated tags. An itemset is hence a feature subset, not a set of goods. For simplifying descriptions, we introduce the following notations:

U	set of customers
F	set of features (folksonomy-based tags)
$T_{i,j}$	indicator of whether jth tag occurred in ith transaction or not
$\sigma(f)$	frequency of occurrence of an itemset $f \subseteq F$ (support count)
$S(f)$	fraction of transactions containing an itemset f (support)
$C(x\|y)$	ratio of co-occurrence of an itemset x over an itemset y (confidence)

Let F be a feature set. A set $I \subseteq F$ is called a large itemset of F if $S(I)$ exceeds maximum support threshold. A set $I \subseteq F$ is called a closed itemset of F if there is no superset $I' \supset I$ such that $I' \subseteq F$.

Large itemsets are often used in association rule mining to catch the key data characteristics which is usually a successful factor of dimension reduction. Due to the label-insensitive nature of large itemsets, the corresponding dimension reduction application is suitable for both supervised and unsupervised learning. In the meanwhile, statistical approaches consider the independently and identically distributed (*i.i.d*) assumption, while closed itemsets can be used to approximately meet the *i.i.d* assumption. It cannot guarantee that generated features are *i.i.d*, but it can reduce the dependence between generated features.

2.2. Multivariate information gain

Let X be the random variable and $p(x)$ is the probability mass function of outcome x. The *Shannon entropy* [15] is defined as

$$H(X) = \sum_{x} -p(x)\log_2 p(x) \tag{1}$$

Let X and Y be two random variables. $p(y)$ and $p(y|x)$ are the probability mass function of outcome y and the conditional probability mass function of outcome y is given as$X = x$, respectively. The *information gain* [14] is defined as

$$IG(Y|X) = H(Y) - \sum_{x} p(x)\left(\sum_{y} -p(y|x)\log_2 p(y|x)\right) \tag{2}$$

The *information gain* is a criterion to evaluate the quality of features. However, the original *information gain* is used to quantify the difference between response and individual feature. Popular multivariate quantify is *multivariate mutual information* measuring the information of a feature set and the response. For the purpose discrimination, only total contribution of feature set to the response is needed to be considered. Other information may distract the original purpose. According to this observation, we propose a new quantity, *multivariate information gain*, measuring the total contribution of feature set to the response for the discrimination purpose.

Definition 1. The *multivariate information gain* of random variables $X_1,\ldots,X_n$ on observation Y is defined as follows:

$$MIG(Y|X_1,\ldots,X_n) = H(Y) - \sum_{x_1,\ldots,x_n} p(x_1,\ldots,x_n) \times \left(-\sum_{y} p(y|x_1,\ldots,x_n)(\log_2 p(y|x_1,\ldots,x_n)\right)$$

where $H(Y)$ is the *Shannon entropy* of random variable Y.

3. Easy-to-explain feature synthesis problem

Original features selected to describe certain phenomena by domain experts are explainable. Nevertheless, attractable phenomena are usually complex and the selection criterion of features for domain experts is based on the relevance between these features and interested phenomenon whether the relevance is significant or not. It causes high dimensional data and feature dependency problem which are dangerous to the performance of learning systems. Traditional dimension reduction techniques search the optimal feature subsets based on different biases. Feature selection algorithms assume that the optimal solution exists in the power set of original feature set, while the assumption of feature extraction algorithms is the existence of optimal solution in a specific feature space generated from original features via certain operators. Due to the complexity of modern data, feature selection algorithms do not perform well with linear systems providing explainable results. It is potentially dangerous for feature extraction algorithms to erase the explanation power of generated features if search space is not restricted in the human explainable space.

Disjunction of conjunctive form is one of the human understandable knowledge form. For example, three features "Idol", "Album" and "Female" can be interpreted as three distinct user preferences, and any disjunction of conjunctive form of these features is still interpretable. ("Idol and Album" or "Idol and Female") can also be interpreted as this user interests in Idol's Album or Female Idol. To maintain the explanation power of generated features by feature extraction algorithms, operators for generating new features should be limited in logic operators. Therefore, new features are explainable because every logic expression can be transformed into disjunction of conjunctive form. We can hence define easy-to-explain features and easy-to-explain feature synthesis problem as follows:

Definition 2. For a feature set $F = \{f_1, f_2, \ldots, f_n\}$, f' is easy-to-explain if f' is a subset of $2^F - \{\emptyset\}$, the power set of F does not contain the empty set.

Easy-to-explain feature synthesis problem definition

Given a training data D and the corresponding feature set F, find a minimal easy-to-explain feature set F' corresponding to F such that *multivariate information gain* of F' over D is maximum.

Each element in the power set of F can be considered as the conjunctive form of original features and a set of these elements is hence a disjunction of conjunctive form. Easy-to-explain feature space is much larger than that of feature selection algorithms search. For given n features, the search space of feature selection algorithms contains only $d=2^n-1$ potential candidates while there are $2^{2^d-1}-1$ candidates in easy-to-explain feature space. The huge number of candidates raises the computational cost for searching optimal solution in acceptable time. Therefore, heuristic approaches can be applied to solve this optimization problem. Our proposed heuristic feature synthesis approach will be discussed in the next section.

4. Technical approach

In e-commerce, social recommendation system has won its reputation. The success of social recommendation system relies on precise social context. Folksonomy-based tags are common description of video contents and usually used to build the user preference profile. These tags can be considered as Boolean-typed features (variables or attributes) in machine learning field. Table 1 shows an example of partial user log matrix for a mobile video order record. Row i represents that a tag subset of predefined tags accompanies the video purchased by a customer in the ith transaction. User's preference can be considered as the disjunctive of the corresponding active tags. For instance, U_5 may prefer Chinese song or fast song or rookie performer according to the last transaction.

However, folksonomy-based tag system which lacks semantic consistency control usually contains feature dependency problem. This user preference model cannot provide precise user's characteristic and hence leads to doubtful social context information from poor fancier clustering results. We aim to search an optimal feature subset to describe users' characteristic as well as to improve the precision of social recommendation system. Different from traditional feature extraction techniques, the generated features should be easy-to-explain.

4.1. Screening low-informative features

According to Table 1, it is easy to find the tags: "Canto" with very low frequency of utilization and "Chinese" appears in all transactions. For the purpose of discriminating users with different preferences, these tags are less relevant to prediction performance, i.e. low-informative. Theorem 1 supports that a feature does not benefit in *multivariate information gain* if almost instances act the same on this feature. This leads to our first heuristic that features with almost the same value in collected data are useless for discrimination. From the viewpoint of dimension reduction, it is acceptable to retain similar discrimination information after eliminating numerous features.

According to Corollary 1, our first heuristic is to evaluate the support of each feature to determine whether the feature is informative or not. If a feature's support value is less than the minimum threshold or more than the maximum threshold, then this feature is useless for characteristic analysis. Corollary 1 also provides an implicit relationship between the discrimination information loss and support threshold while the support count of screened tag approaches to 0 or total number of instances. The proof of Theorem 1 is provided in Appendix.

Theorem 1. *Let $Y,F_1,\ldots,F_n$ be random variables and m be the number of instances, if random variable F_n has an outcome a, then* $\lim_{P(f_n=a)\to 1} MIG(Y|F_1,\ldots,F_n) = MIG(Y|F_1,\ldots,F_{n-1})$.

Corollary 1. *Let $Y,F_1,\ldots,F_n$ be random variables and m be the number of instances, if for each i, F_i has binary outcomes, then* $\lim_{\sigma(f_n)\to m}\mathrm{MIG}(Y|F_1,\ldots,F_n)=\mathrm{MIG}(Y|F_1,\ldots,F_{n-1})$ *and* $\lim_{\sigma(f_n)\to 0}\mathrm{MIG}(Y|F_1,\ldots,F_n)=\mathrm{MIG}(Y|F_1,\ldots,F_{n-1})$.

Proof. Because F_n has binary outcomes and $\sigma(f_n)$ approaches to m, it implies that $P(F_n=1)$ approaches to 1. By Theorem 1, we can achieve that $\lim_{\sigma(f_n)\to m}\mathrm{MIG}(Y|F_1,\ldots,F_n)=\mathrm{MIG}(Y|F_1,\ldots,F_{n-1})$. Similarly, we can also achieve that $\lim_{\sigma(f_n)\to 0}\mathrm{MIG}(Y|F_1,\ldots,F_n)=\mathrm{MIG}(Y|F_1,\ldots,F_{n-1})$ by Theorem 1, because $\sigma(f_n)\to 0 \Rightarrow P(F_n=0)\to 1$.

4.2. Shrinking high-dependent features

From the view of social recommendation systems, the characteristics of users' preference are usually assumed as disjunctive of original features. Users' preferences should be characterized more precisely instead of disjunctive form, while high dependent tags should be shrunk into one feature for reducing over-concerning. For example, "slow songs" and "lyric" are always co-occurrence in Table 1. The new feature "slow songs and lyric" will more precisely describe users' preference, instead of "slow songs" and "lyric". From the theoretic views, these two tags contribute similar discrimination information and therefore are high-redundant. Theorem 2 guarantees that shrinking high-dependent (High-redundant) features affect *multivariate information gain* slightly if the distributions of these two random variables (features) are almost the same. This leads to our second heuristic which merging two high-dependent features into a new feature does not harm the prediction performance much. The proof of Theorem 2 is provided in Appendix.

Table 1
A partial original user log.

No.	User	Features										
		Slow songs	Lyric	Single	Male	Canto	Chinese	Fast songs	Idol	Crude	Hot	rookie
1	U_1	1	1	1	0	0	1	0	0	1	0	0
2	U_2	1	1	0	0	0	1	1	0	1	1	0
3	U_3	1	1	0	0	0	1	1	0	1	1	0
4	U_2	0	0	0	0	1	1	0	0	1	0	1
5	U_1	1	1	1	1	0	1	0	0	1	0	0
6	U_4	1	1	1	0	0	1	0	0	1	0	0
7	U_4	1	1	0	0	0	1	1	0	0	0	0
8	U_5	0	0	1	0	0	1	0	0	1	0	1
9	U_6	1	1	0	1	0	1	1	0	1	1	0
10	U_5	0	0	0	0	0	1	1	0	0	0	1

Theorem 2. *Let $Y,F_1,\ldots,F_n$ be random variables and m be the number of instances, if for all $\varepsilon>0$, there exists a $\delta>0$, $|P(f_{n-1}=f_n)-1|<\delta$ such that $|\mathrm{MIG}(Y|F_1,\ldots,F_n)-\mathrm{MIG}(Y|F_1,\ldots,F_{n-2},\{F_{n-1},F_n\})|<\varepsilon$.*

Corollary 2. *Let $Y,F_1,\ldots,F_n$ be random variables and m be the number of instances, if for each i, F_i has binary outcomes, then $\lim_{C(f_{n-1}|f_n)\to 1}\mathrm{MIG}(Y|F_1,\ldots,F_n)=\mathrm{MIG}(Y|F_1,\ldots,F_{n-2},\{F_{n-1},F_n\})$.*

4.3. Enhancing dominated features (optional)

In entertainment video business, a user purchasing a video indicates that the property of this video meets his preference. However, there is no obvious information when user does not purchase some video contents. That is, feature does not contribute equally on different feature value in recommending videos. The only information of this tag provides for discrimination if it appears in purchase history. Following this observation, feature with higher frequency in transactions will bring more information about users' preferences. From the viewpoint of association rule mining, large itemsets represent the significant characteristics of data. Therefore, our last heuristic is to enhance the features with obvious information by mining the large itemsets.

4.3.1. Feature synthesis algorithm

Input:
Transactions Data $[C|T]_{m\times(|F|+1)}$ where C records which customer in each transaction, T is the tag-transaction mapping matrix and F is the set of tags.
Maximum support threshold, T_M
Minimum support threshold, T_m
Minimum support threshold, T_m',
Confidence threshold, T_c
Output: easy-to-explain feature subset F
Method:
STEP 1. Build the user profile $P_{|U|\times|F|}$ by calculating the frequency of features in Transaction Data $[C|T]$.
STEP 2. Normalize each column vector $P(:,i)$.
STEP 3. for each feature f in F, remove f from F if $S(f)>T_M$ or $S(f)<T_m$.
STEP 4. for each feature f_i in F, find each feature f_j in F with $P(f_i,f_j)>T_c$, remove f_j from F,add $f_i\cup f_j$ into F and $P(:,I(f_i\cup f_j))=P(:,i)+P(:,j)$ where $I(f_i\cup f_j)$ is the index of new feature $f_i\cup f_j$ in F. Remove f_i from F if there exists a feature containing f_i.
STEP 5. Repeat **STEP 4** until the cardinality of F remains the same.
STEP 6. for each feature f, remove f from F if $S(f)<T_{m'}$. (optional)
The proposed feature synthesis algorithm follows the above heuristics to search easy-to-explain features for maximizing *multivariate information gain* and hence to improve the precision of social recommendation systems. Step 3 screens the low-informative features. Steps 4 and 5 shrink the high-dependent features iteratively. This shrinking phase also transforms original Boolean-typed features into discrete-typed features for redeeming the potential information loss. If the confidence of two features approximates 1, some customers may be interested in only one feature. This fuzzy recovery will retain this kind of preference differences. Notice that there is only one feature used to merge with other features in each round. Therefore, after each round, the number of features is equal to the previous one or decreased by 1, relying on the selected feature highly depending on some of the rest features or not. Step 6 enhances the dominated features which bring more information of describing users' preferences. The enhancement we adopt here is the extreme Boolean version which is to keep dominated features and to remove the rest. Theorem 3 guarantees that feature synthesis algorithm produces easy-to-explain features. Our proposed feature synthesis algorithm search part of easy-to-explain features which are the conjunctive form of original features. It is because that the applying linear function (similarity measurement) will lead to simultaneous interpretation of whole features. However, this search space is still larger than that of feature selection algorithms search ($2^{2^n-1}-1$ compared to 2^n-1).

Theorem 3. *Feature synthesis algorithm produces easy-to-explain features.*

Proof. It can easily be verified that all steps of feature synthesis algorithm on feature set are logic operators. Because every logic expression can be transformed into a disjunction of conjunctive form, the produced features are in the disjunction of conjunctive form of original features. Therefore, we can claim that feature synthesis algorithm produces easy-to-explain features.

5. Experiments

5.1. Experimental design

In this paper, we aim to search an easy-to-explain feature subset which can properly characterize the users' preferences while conquering the curse of dimensionality. The experiment data were offered by the largest commercial entertainment video Web shop in Taiwan. There are total 1487 available videos and purchase information of 10,906 customers during 2008/06/12–2009/05/07. Each video contains several suitable tags selected from 117 predefined tags. The 28,249 transactions are separated into training data (from 2008/06/12 to 2008/11/23) and testing data (from 2008/11/24 to 2009/05/07), The training data is further split into training set and validation set for the purpose of model selection. Fig. 1 shows the data sizes and the corresponding duration and the detail of data statistics is shown in Table 2.

Our feature synthesis algorithm consists of three phases: feature screening phase, feature shirking phase and feature enhancing phase. In feature screening phase, only features with support between 0.1 (T_m) and 0.9 (T_M) are kept. While shirking features with confidence over 0.9 (T_c) in feature shirking phase, only features with support over 0.5 ($T_{m'}$) are kept in feature enhancing phase.

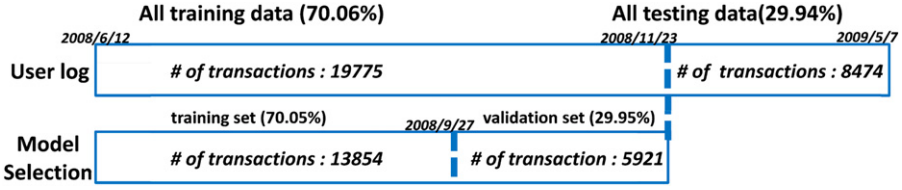

Fig. 1. The data set of all user log and Model Selection.

Content-based collaborative filtering (CBCF) recommendation approach is based on the assumption that people will be interested in those persons with similar preferences in interests. The well-known k-means clustering algorithm [7] is used to cluster users with similar preferences. The similarity measures, cosine similarity, are selected for k-means because of discrete type of feature. In particularly, for two vectors V_1 and V_2, the cosine similarity of V_1 and V_2, $Cos(V_1,V_2)=V_1\cdot V_2/|V_1||V_2|$. The cluster number for k-means algorithm is selected from 2 to 10 via model selection mechanism. The lengths of recommendation lists are set as 5 and 10. The experimental results will be analyzed and discussed in the next Section.

5.2. Experimental results

The goal of recommendation system is to recommend products which customers will be interested in and purchase. The hit rate is a suitable criterion to measure the performance of a recommendation system. We call hit if one of someone's testing data is in the recommendation list. So the hit rate means the fraction of hit number divided by the number of total testing users. Our recommendation is based on CBCF with k-means clustering algorithm. However, k-means algorithm is very sensitive on initial setting. The first randomly selected k centroids will affect the performance of k-means algorithm much. To avoid this issue of k-means, we repeat our experiments 50 times on each attribute set. For each experiment, the number of clusters is decided by model selection.

5.2.1. Effectiveness evaluation

The experimental result is shown in Table 3. The results show that our searched feature subsets perform well in recommending entertainment videos, especially the feature screening phase improves the most hit rate. This explains that low informative features have bad influences on recommending video and this serious problem exists in folksonomy-based tag system. The feature enhancing phase improves performance when the recommendation list length is 5. This is because little recommendations need to focus on the sufficient users' preferences. When producing 10 recommendations, the feature enhancing phase does not perform well since producing long recommendation list requires more information which is ignored by this phase.

For comparison purpose, the well-known wrapper-based feature selection techniques, forward feature selection (FFS) technique and backward feature elimination (BFE), have been selected for comparison. FFS evaluates each candidate feature subset and selects the best candidate feature subset to generate the next candidate feature subsets by adding one remain feature into it. FFS stops if the new selected feature does not benefit in prediction performance. On the contrast, the candidate feature subsets of BFE are formed by eliminating one feature from current best feature subsets. In our experiments, each candidate feature subset is evaluated by the same model selection procedure which is decided by different cluster number setting for k-means algorithm (from 2 to 10). We also repeat 50 times our experiments on feature subsets selected by FFS and BFE. The comparison is also shown in Table 3. FFS performs poorly (average hit rates are 3.35% (the length of recommendation list $L=5$) and 7.059% ($L=10$)), while BFE performs slightly better than FFS (5.455% ($L=5$) and 11.214% ($L=10$)). The results show that our proposed feature synthesis algorithm outperforms FFS and BFE on this data set.

Best k in model selection is also a criterion to evaluate the quality of clustering algorithms. Cluster number refers to the degree of discrimination between distinct customers' preferences. Table 3 shows that Best k increases when each phase of our feature synthesis algorithm applies. For example, the characteristic "Idol and Album" and "Idol and Female" are different clusters' characteristic. According to our observations, "Idol" is the characteristic of single cluster; we can use the more precise features to describe the users' preferences. This leads to more precisely fancier clustering results as well as more successful recommendations. Our experimental results demonstrate that out proposed feature synthesis algorithm performs well on the entertainment video recommendation in three aspects: (1) better characterization of users' preferences, (2) easier interpretation of clustering results and (3) better performance of recommendation system.

5.2.2. Process time comparison

The experiments are run on an IBM eServer x3400 server with two Intel® Xeon® Processor E5420 (12 M Cache, 2.50 GHz and 1333 MHz FSB), 4 Gb RAM and Microsoft Windows Server 2008 SP2 operating system. The algorithm is implemented by Microsoft Visual Studio 2008 and Microsoft SQL server 2005. We further apply parallel programming technique to speed up the exhaustive computations of forward feature selection (FFS) technique and backward feature elimination (BFE) technique. Our proposed feature synthesis algorithm is a kind of filter-based feature selection techniques. Therefore, it produces the same feature subset regardless of the recommendation list length. In contrast, FFS and BFE are both sensitive on the

Table 2
The details of data statistics of purchase information.

		Transaction number (percentage)	Customer number (percentage)	Avg. purchase per customer (std.)
Training data	Training set	13854 (49.04%)	5869 (53.81%)	2.36 (1.87)
	Validation set	5921 (20.96%)	2640 (24.21%)	2.43 (4.08)
	Total	19775 (70.06%)	8509 (78.02%)	2.32 (3.71)
Testing data		8474 (29.94%)	3454 (31.67%)	

Table 3
The results of recommendations.

Feature set	List length (L): 5		List length (L): 10		Feature number	
	Best K	Avg. hit rate (Std.)	Best K	Avg. hit rate (Std.)		
1-itemset (without screening)	3	4.080% (0.013)	5	10.885% (0.029)	117	
1-itemset (with screening)	6	6.114% (0.014)	7	11.570% (0.015)	44	
Closed itemset (with screening)	8	6.930% (0.008)	8	11.954% (0.025)	17	
Large closed itemset (with screening)	8	7.262% (0.008)	7	11.598% (0.018)	15	
Forward feature selection	9	3.35% (0.009)	8	7.059% (0.013)	$L=5$	$L=10$
					5	6
Backward feature elimination	9	5.455% (0.023)	10	11.214% (0.106)	$L=5$	$L=10$
					114	115

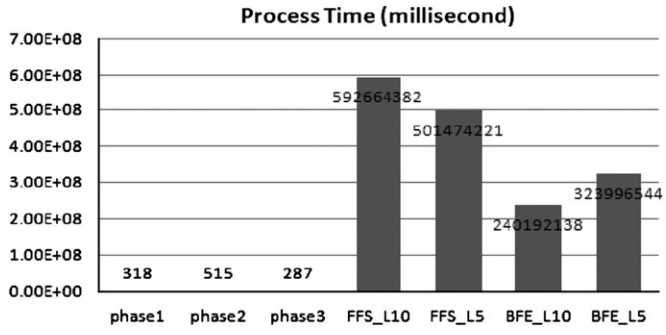

Fig. 2. The Process Time of dimension reduction algorithms.

setting of the recommendation list length (L). The comparison result is shown in Fig. 2. The x-axis represents different dimension reduction approaches, including three phases of our approach (phase 1, phase 2 and phase 3), FFS (FFS_L5 and FFS_L10) and BFE (BFE_L5 and BFE _L10) with different recommendation list length setting. The y-axis shows process time in millisecond.

6. Conclusion

For the purpose of providing explainable predicted results and improving the performance of learning system, we define easy-to-explain features to provide large search space for feature extraction algorithm. Upon easy-to-explain search space, linear learning system can achieve better performance while the predicted results are still explainable. Our proposed feature synthesis algorithm is unsupervised and lightweight. We described the algorithm in the context of clustering and social recommendation, but we suppose its applicability can be much broader. The theoretical constraints are relaxed to achieve practical benefits. This relaxation motivates future research attentions on exploring the relations among joint mass distribution of features, support and confidence. The experimental results show that our approach can obtain the interpretable clustering results as well as improve the recommendation.

Acknowledgment

This work was partially supported by National Science Council of the Republic of China under contracts NSC97-2511-S-468-004-MY3 and NSC 98–2851-S-468-004-MY3.

Appendix. Proofs of Theorems 1 and 2

Multivariate information gain measures the quantity of uncertainty change of interested phenomenon after giving several features. It can be considered as the amount of discrimination information from these features. In Theorems 1 and 2, we are interested in the influence of individual feature on discrimination information, especially in features with support approximating to 1 or 0 and feature sets with confidence approximating to 1. To achieving these goals, we find upper bounds and lower bounds of *multivariate information gain*. Theorems 1 and 2 state that these upper bounds and lower bounds approach identical in different cases.

Lemma 1. *Let $F_1,\ldots,F_n$ be random variables and m be the number of instances, if there are k instances with feature $f_n = a$ then*

$$p(f_1,\ldots,f_{n-1}) - \frac{m-k}{m} \le p(f_1,\ldots,f_{n-1},f_n = a) \le p(f_1,\ldots,f_{n-1})$$

and

$$p(f_1,\ldots,f_{n-1}) - \frac{k}{m} \le p(f_1,\ldots,f_{n-1},f_n \ne a) \le p(f_1,\ldots,f_{n-1}).$$

Proof. Because of the definition of joint probability, we have

$$p(f_1,\ldots,f_{n-1}) = p(f_1,\ldots,f_{n-1},f_n = a) + p(f_1,\ldots,f_{n-1},f_n \ne a), \quad (3)$$

and also simple inequalities that

$$0 \le p(f_1,\ldots,f_{n-1},f_n \ne a) \le p(f_n \ne a) = \frac{m-k}{m}. \quad (4)$$

By (3) and (4), it can be derived that

$$p(f_1,\ldots,f_{n-1}) - \frac{m-k}{m} \le p(f_1,\ldots,f_{n-1},f_n = a) \le p(f_1,\ldots,f_{n-1}). \quad (5)$$

Similarly, $0 \le p(f_1,\ldots,f_{n-1},f_n = a) \le p(f_n = a) = k/m$. We can also derive that

$$p(f_1,\ldots,f_{n-1}) - \frac{k}{m} \le p(f_1,\ldots,f_{n-1},f_n \ne a) \le p(f_1,\ldots,f_{n-1}). \quad \square \quad (6)$$

Proof of Theorem 1. By Definition 1, we have

$$MIG(Y|F_1,\ldots,F_n) = H(Y) - \sum_{f_1,\ldots,f_n} p(f_1,\ldots,f_n)\left(-\sum_y p(y|f_1,\ldots,f_n)\log_2 p(y|f_1,\ldots,f_n)\right). \quad (7)$$

When given different feature sets, only the second term of right hand-side of eq. (7) varies which is denoted by $E_{F_1,\ldots,F_n}(H(Y|F_1,\ldots,F_n))$.

$$\begin{aligned}
&\sum_{f_1,\ldots,f_n} p(f_1,\ldots,f_n)\left(-\sum_y p(y|f_1,\ldots,f_n)\log_2 p(y|f_1,\ldots,f_n)\right)\\
&= \sum_{f_1,\ldots,f_{n-1}} p(f_1,\ldots,f_{n-1},f_n = a)\\
&\left(-\sum_y p(y|f_1,\ldots,f_{n-1},f_n = a)\log_2 p(y|f_1,\ldots,f_{n-1},f_n = a)\right)\\
&+ \sum_{f_1,\ldots,f_{n-1}} p(f_1,\ldots,f_{n-1},f_n \ne a)\\
&\left(-\sum_y p(y|f_1,\ldots,f_{n-1},f_n \ne a)\log_2 p(y|f_1,\ldots,f_{n-1},f_n \ne a)\right). \quad (8)
\end{aligned}$$

By Lemma 1 and the definition of conditional probability,

$$\begin{aligned}
&\sum_{f_1,\ldots,f_{n-1}} p(f_1,\ldots,f_{n-1},f_n = a)\\
&\left(-\sum_y p(y|f_1,\ldots,f_{n-1},f_n = a)\log_2 p(y|f_1,\ldots,f_{n-1},f_n = a)\right)\\
&\le \sum_{f_1,\ldots,f_{n-1}} p(f_1,\ldots,f_{n-1})\\
&\left(-\sum_y \frac{p(y,f_1,\ldots,f_{n-1},f_n = a)}{P(f_1,\ldots,f_{n-1}f_n = a)}\log_2 \frac{p(y,f_1,\ldots,f_{n-1},f_n = a)}{P(f_1,\ldots,f_{n-1},f_n = a)}\right)\\
&\le \sum_{f_1,\ldots,f_{n-1}} p(f_1,\ldots,f_{n-1})\left(-\sum_y \frac{p(y,f_1,\ldots,f_{n-1})}{P(f_1,\ldots,f_{n-1}) - (m-k/m)}\right.
\end{aligned}$$

$$\log_2 \frac{p(y,f_1,\ldots,f_{n-1})-((m-k)/m)}{P(f_1,\ldots,f_{n-1})}\Bigg)$$

Since $p(f_1,\ldots,f_{n-1},f_n \neq a)$ approaches to zero as the number of instances with feature $f_n = a$ approaches to m, we have

$$\lim_{P(f_n=a)\to 1} E_{F_1,\ldots,F_n}(H(Y|F_1,\ldots,F_n)) \le E_{F_1,\ldots,F_{n-1}}(H(Y|F_1,\ldots,F_{n-1})). \tag{9}$$

We can also apply Lemma 1 to derive the lower bound.

$$\sum_{f_1,\ldots,f_{n-1}} p(f_1,\ldots,f_{n-1},f_n=a)$$
$$\left(-\sum_y p(y|f_1,\ldots,f_{n-1},f_n=a)\log_2 p(y|f_1,\ldots,f_{n-1},f_n=a)\right)$$
$$\ge \sum_{f_1,\ldots,f_{n-1}} \left(p(f_1,\ldots,f_{n-1})-\frac{m-k}{m}\right)$$
$$\left(-\sum_y \frac{p(y,f_1,\ldots,f_{n-1},f_n=a)}{P(f_1,\ldots,f_{n-1},f_n=a)} \log_2 \frac{p(y,f_1,\ldots,f_{n-1},f_n=a)}{P(f_1,\ldots,f_{n-1},f_n=a)}\right)$$
$$\ge \sum_{f_1,\ldots,f_{n-1}} \left(p(f_1,\ldots,f_{n-1})-\frac{m-k}{m}\right)$$
$$\left(-\sum_y \frac{p(y,f_1,\ldots,f_{n-1})-m-k/m}{P(f_1,\ldots,f_{n-1})} \log_2 \frac{p(y,f_1,\ldots,f_{n-1})}{P(f_1,\ldots,f_{n-1})-m-k/m}\right)$$

When the number of instances with feature $f_n = a$ approaches to m, we can derive that

$$\lim_{P(f_n=a)\to 1} E_{F_1,\ldots,F_n}(H(Y|F_1,\ldots,F_n)) \ge E_{F_1,\ldots,F_{n-1}}(H(Y|F_1,\ldots,F_{n-1})). \tag{10}$$

According to inequalities (9) and (10), by the sandwich theorem, we can claim that

$$\lim_{P(f_n=a)\to 1} E_{F_1,\ldots,F_n}(H(Y|F_1,\ldots,F_n)) = E_{F_1,\ldots,F_{n-1}}(H(Y|F_1,\ldots,F_{n-1}))$$

Therefore, $\lim_{P(f_n=a)\to 1}\mathrm{MIG}(Y|F_1,\ldots,F_n) = \mathrm{MIG}(Y|F_1,\ldots,F_{n-1})$. Based on Lemma 1 and the similar idea, we can also derive that $\lim_{P(f_n=a)\to 1}\mathrm{MIG}(Y|F_1,\ldots,F_n) = \mathrm{MIG}(Y|F_1,\ldots,F_{n-1})$

Lemma 2. Let $F_1,\ldots,F_n$ be random variables and m be the number of instances, if there are k instances with features $f_{n-1} = f_n$ then

$$p(f_1,\ldots,f_{n-1})-p(f_{n-1},f_{n-1}^c) \le p(f_1,\ldots,f_{n-1},f_{n-1}) \le p(f_1,\ldots,f_{n-1}).$$

Proof. Because of the definition of joint probability, we have

$$p(f_1,\ldots,f_{n-1}) = p(f_1,\ldots,f_{n-1},f_{n-1})+p(f_1,\ldots,f_{n-1},f_{n-1}^c), \tag{11}$$

and also simple inequalities that

$$0 \le p(f_1,\ldots,f_{n-1},f_{n-1}^c) \le p(f_{n-1},f_{n-1}^c), \tag{12}$$

where f_{n-1}^c is the complement of f_{n-1}. By (11) and (12), it can be derived that

$$p(y,f_1,\ldots,f_{n-1})-p(f_{n-1},f_{n-1}^c) \le p(y,f_1,\ldots,f_{n-1},f_{n-1}) \le p(y,f_1,\ldots,f_{n-1}). \tag{13}$$

Proof of Theorem 2. By Definition 1, we have that

$$MIG(Y|F_1,\ldots,F_n) = H(Y)-\sum_{f_1,\ldots,f_n} p(f_1,\ldots,f_n)$$
$$\left(-\sum_y p(y|f_1,\ldots,f_n)\log_2 p(y|f_1,\ldots,f_n)\right).$$

When given different feature sets, only the second term of right hand-side of eq. (7) varies which is denoted by $E_{F_1,\ldots,F_n}(H(Y|F_1,\ldots,F_n))$.

$$\sum_{f_1,\ldots,f_n} p(f_1,\ldots,f_n)\left(-\sum_y p(y|f_1,\ldots,f_n)\log_2 p(y|f_1,\ldots,f_n)\right)$$
$$= \sum_{f_1,\ldots,f_{n-1}} p(f_1,\ldots,f_{n-1},f_{n-1})$$
$$\left(-\sum_y p(y|f_1,\ldots,f_{n-1},f_{n-1})\log_2 p(y|f_1,\ldots,f_{n-1},f_{n-1})\right)$$
$$+ \sum_{f_1,\ldots,f_{n-1}} p(f_1,\ldots,f_{n-1},f_{n-1}^c)$$
$$\left(-\sum_y p(y|f_1,\ldots,f_{n-1},f_{n-1}^c)\log_2 p(y|f_1,\ldots,f_{n-1},f_{n-1}^c)\right)$$

By Lemma 2 and the definition of conditional probability,

$$\sum_{f_1,\ldots,f_{n-1}} p(f_1,\ldots,f_{n-1},f_{n-1})$$
$$(-\sum_y p(y|f_1,\ldots,f_{n-1},f_{n-1})\log_2 p(y|f_1,\ldots,f_{n-1},f_{n-1}))$$
$$\le \sum_{f_1,\ldots,f_{n-1}} p(f_1,\ldots,f_{n-1})$$
$$\left(-\sum_y \frac{p(y,f_1,\ldots,f_{n-1},f_{n-1})}{P(f_1,\ldots,f_{n-1},f_{n-1})} \log_2 \frac{p(y,f_1,\ldots,f_{n-1},f_{n-1})}{P(f_1,\ldots,f_{n-1},f_{n-1})}\right)$$

by (13)

$$\le \sum_{f_1,\ldots,f_{n-1}} p(f_1,\ldots,f_{n-1})$$
$$\left(-\sum_y \frac{p(y,f_1,\ldots,f_{n-1})}{P(f_1,\ldots,f_{n-1})-p(f_{n-1},f_{n-1}^c)} \log_2 \frac{p(y,f_1,\ldots,f_{n-1})-p(f_{n-1},f_{n-1}^c)}{P(f_1,\ldots,f_{n-1})}\right)$$

(By Lemma 2)

Since $p(f_{n-1},f_{n-1}^c)$ approaches to zero as $C(f_{n-1}|f_n)$ approaches to 1, we have

$$\lim_{C(f_{n-1}|f_n)\to 1} E_{F_1,\ldots,F_n}(H(Y|F_1,\ldots,F_n)) \le E_{F_1,\ldots,F_{n-1}}(H(Y|F_1,\ldots,F_{n-1})). \tag{14}$$

Again, we apply Lemma 2 to achieve lower bound.

$$\sum_{f_1,\ldots,f_{n-1}} p(f_1,\ldots,f_{n-1},f_{n-1})$$
$$\left\{-\sum_y p(y|f_1,\ldots,f_{n-1},f_{n-1})\log_2 p(y|f_1,\ldots,f_{n-1},f_{n-1})\right\}$$
$$\ge \sum_{f_1,\ldots,f_{n-1}} (p(f_1,\ldots,f_{n-1})-p(f_{n-1},f_{n-1}^c))$$
$$\left(-\sum_y \frac{p(y,f_1,\ldots,f_{n-1},f_{n-1})}{P(f_1,\ldots,f_{n-1},f_{n-1})} \log_2 \frac{p(y,f_1,\ldots,f_{n-1},f_{n-1})}{P(f_1,\ldots,f_{n-1},f_{n-1})}\right)$$
$$\ge \sum_{f_1,\ldots,f_{n-1}} (p(f_1,\ldots,f_{n-1})-p(f_{n-1},f_{n-1}^c))$$
$$\left(-\sum_y \frac{p(y,f_1,\ldots,f_{n-1},f_{n-1})-p(f_{n-1},f_{n-1}^c)}{P(f_1,\ldots,f_{n-1},f_{n-1})}\right.$$
$$\left.\log_2 \frac{p(y,f_1,\ldots,f_{n-1},f_{n-1})}{P(f_1,\ldots,f_{n-1},f_{n-1})-p(f_{n-1},f_{n-1}^c)}\right).$$

When $C(f_{n-1}|f_n)$ approaches to 1, we can derive that

$$\lim_{C(f_{n-1}|f_n)\to 1} E_{F_1,\ldots,F_n}(H(Y|F_1,\ldots,F_n)) \geq E_{F_1,\ldots,F_{n-1}}(H(Y|F_1,\ldots,F_{n-1})). \quad (15)$$

According to inequalities (14) and (15), we can claim that

$$\lim_{C(f_{n-1}|f_n)\to 1} E_{F_1,\ldots,F_n}(H(Y|F_1,\ldots,F_n)) = E_{F_1,\ldots,F_{n-1}}(H(Y|F_1,\ldots,F_{n-1}))$$

Let a random variable $\{F_{n-1},F_n\} = F_{n-1}$, then

$$\lim_{C(f_{n-1}|f_n)\to 1} \mathrm{MIG}(Y|F_1,\ldots,F_n) = \mathrm{MIG}(Y|F_1,\ldots,F_{n-2},\{F_{n-1},F_n\})$$

Reference

[1] R. Bellman, Adaptive Control Processes: A Guided Tour, Princeton University Press, 1961.
[2] M. Cheng, B. Fang, C.M. Pun, Y.Y. Tang, Kernel-view based discriminant approach for embedded feature extraction in high-dimensional space, Neurocomputing 74 (9) (2011) 1478–1484.
[3] W.S. Chu, J.C. Chen, J.J.J. Lien, Kernel discriminant transformation for image set-based face recognition, Pattern Recognition 44 (8) (2011) 1567–1580.
[4] P. Comon, Independent Component Analysis: a new concept? Signal Process. 36 (3) (1994) 287–314.
[5] F. Hussein, N. Kharma, R. Ward, Genetic algorithms for feature selection and weighting, a review and study, Proc. sixth Int. Conf. Doc. Anal. Recognit. (2001) 1240–1244.
[6] I.T. Jollife, Principal Component Analysis, first Ed., Springer, New York, 1986.
[7] T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, A.Y. Wu, An efficient *k*-means clustering algorithm: analysis and implementation, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 881–892.
[8] C. Lai, M.J.T. Reinders, L. Wessels, Random subspace method for multivariate FS, Pattern Recognit. Lett. 27 (10) (2006) 1067–1076.
[9] K.C. Li, Sliced inverse regression for dimension reduction (with discussion), J. Am. Stat. Assoc. 86 (1991) 316–342.
[10] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, first ed., Academic Press, 1995.
[11] J. McBain, M. Timusk, Feature extraction for novelty detection as applied to fault detection in machinery, Pattern Recognit. Lett. 32 (7) (2011) 1054–1067.
[12] S. Nakariyakul, D.P. Casasent, Adaptive branch and bound algorithm for optimal FS, Pattern Recognit. Lett. 28 (12) (2007) 1415–1427.
[13] F.M. Nejad, H. Zakeri, An optimum feature extraction method based on Wavelet-Radon transform and Dynamic Neural Network for pavement distress classification, Expert Syst. Appl. 38 (2011) 9442–9460.
[14] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.
[15] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (379–423) (1948) 623–656.
[16] P. Somol, P. Pudil, Oscillating search algorithms for feature selection, Proc. 15th Int. Conf. Pattern Recognit. 2 (2000) 406–409.
[17] P. Somol, P. Pudil, J. Kittler, Fast branch & bound algorithms for optimal feature selection, IEEE Trans. Pattern Anal. Mach. Intell. 26 (7) (2004) 900–912.
[18] Y. Sun, S. Todorovic, S. Goodison, Local-learning-based feature selection for high-dimensional data dnalysis, IEEE Trans. Pattern Anal. Mach. Intell 32 (9) (2010) 1610–1626.
[19] M.A. Tahir, A. Bouridane, F. Kurugollu, Simultaneous feature selection and feature weighting using hybrid tabu search/k-nearest neighbor classifier, Pattern Recognit. Lett. 28 (4) (2007) 438–446.
[20] H.M. Wu, Kernel sliced inverse regression with applications on classification, J. Comput. Graph. Stat. 17 (3) (2008) 590–610.
[21] F. Yang, K.Z. Mao, Robust feature selection for microarray data based on multicriterion fusion, IEEE/ACM Trans. Comput. Biol. Bioinform. 8 (4) (2011) 1080–1092.
[22] W.K. Yang, C.Y. Sun, L. Zhang, A multi-manifold discriminate analysis method for image feature extraction, Pattern Recognition 44 (8) (2011) 1649–1657.
[23] Y.R. Yeh, S.Y. Huang, Y.J. Lee, Nonlinear dimension reduction with Kernel sliced inverse regression, IEEE Trans. Knowl. Data Eng. 21 (11) (2009) 1590–1603.
[24] Z. Zhu, Y. Ong, M. Dash, Wrapper-filter feature selection algorithm using a memetic framework, IEEE Trans. Syst. Man Cybern. B Cybern. 37 (1) (2007) 70–76.

Shian-Shyong Tseng received the Ph.D. degree in computer engineering from the National Chiao Tung University in 1984. From 1983 to 2009, he was on the faculty of the Department of Computer and Information Science at National Chiao-Tung University. From 1991 to 1992 and 1996 to 1998, he acted as the Chairman of Department of Computer and Information Science. From 1992 to 1996, he was the Director of the Computer Center at Ministry of Education and the Chairman of Taiwan Academic Network (TANet) management committee. In Dec. 1999, he founded Taiwan Network Information Center (TWNIC) and was the Chairman of the board of directors of TWNIC from 1999 to 2005. He was the Dean of the College of Computer Science, Asia University from 2005 to 2008. He is currently a vice president of ASIA University and the Chairman of the board of directors of TWNIC. Dr. Tseng is an Editor-in-Chief of International Journal of Digital Learning Technology, an editor of International Journal of Fuzzy Systems, Journal of Internet Technology and International Journal of Computational Science, and a member of IEEE and Phi Tau Phi Societies. He is also a Co Editor-in-Chief of Asian Journal of Health and Information Science. He was named an Outstanding Talent of Information Science of the Republic of China in 1989. He obtained the 1992, 1994, and 1995 Outstanding Research Awards of the National Science Council of the Republic of China. He was the winner of the 1990, 1991, 1998 and 2000 Acer Long Term Awards for outstanding M.S. Thesis Supervision and the winner of 1992 and 1996 Acer Long Term Awards for outstanding Ph.D. Dissertation Supervision. He was also awarded Outstanding Youth Honor of R.O.C. in 1992. His current research interests include expert systems, data mining, computer-assisted learning, and Internet-based applications. He has published more than 100 journal papers.

Tsung-Ju Lee received a BS degree in Mathematics from TungHai University, Taiwan in 2000 and an MS degree in Applied Mathematics from National Chiao Tung University, Taiwan in 2002. Currently, he is working towards the Ph.D. degree in the Department of Computer Science, National Chiao Tung University, Taiwan. His current research interests include machine learning, data mining and various applications, especially in network security, e-learning and software testing.