Pattern Recognition Letters 40 (2014) 104–112

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

# Integration of dense subgraph finding with feature clustering for unsupervised feature selection ☆

Sanghamitra Bandyopadhyay [a,*], Tapas Bhadra [a], Pabitra Mitra [b], Ujjwal Maulik [c]

[a] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India
[b] Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur 721302, India
[c] Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032, India

ARTICLE INFO

*Article history:*
Received 15 May 2013
Available online 15 December 2013

*Keywords:*
Pattern recognition
Unsupervised feature selection
Mutual information
Normalized mutual information

ABSTRACT

In this article a dense subgraph finding approach is adopted for the unsupervised feature selection problem. The feature set of a data is mapped to a graph representation with individual features constituting the vertex set and inter-feature mutual information denoting the edge weights. Feature selection is performed in a two-phase approach where the densest subgraph is first obtained so that the features are maximally non-redundant among each other. Finally, in the second stage, feature clustering around the non-redundant features is performed to produce the reduced feature set. An approximation algorithm is used for the densest subgraph finding. Empirically, the proposed approach is found to be competitive with several state of art unsupervised feature selection algorithms.



## 1. Introduction

Over the past decade pattern recognition techniques have been extensively used to solve several real-life problems that involve very high dimensional data. Dimensionality reduction is almost always necessary to remove the redundant features while retaining the salient characteristics of the data as far as possible (Kwak and Choi, 2002).

Feature selection algorithms can be divided into two categories based on the feature evaluation methodology, namely, filter and wrapper methods (Dash and Liu, 1997). In the filter approaches, a candidate feature subset is evaluated at each iteration based on certain statistical measures. Some known filter type approaches are based on *t*-test (Hua et al., 2008), chi-square test (Jin et al., 2006), Wilcoxon Mann–Whitney test (Liao et al., 2007), mutual information (Battiti, 1994; Kwak and Choi, 2002; Peng et al., 2005; Estévez et al., 2009; Vinh et al., 2010), Pearson correlation coefficients (Biesiada and Duch, 2008), etc. On the other hand, wrapper methods utilize the performance of a classifier as the evaluation criteria for measuring the goodness of a candidate feature subset (Kohavi and John, 1997).

Based on the availability of class labels, feature selection algorithms can also be classified in two ways, namely, supervised and unsupervised feature selection. Supervised feature selection is generally employed when the class information are in hand, otherwise unsupervised approach is used. Most known filter type approaches, belong to the category of supervised learning. On the other hand, a limited number of researches have been conducted in the field of unsupervised feature selection. Unsupervised feature selection using feature similarity measure (FSFS) (Mitra et al., 2002), Laplacian Score for Feature Selection (LSFS) (He et al., 2005), SPectral Feature Selection (SPFS) (Zhao and Liu, 2007), Multi Cluster Feature Selection (MCFS) (Cai et al., 2010), Unsupervised Discriminative Feature Selection (UDFS) (Yang et al., 2011), etc. are some existing algorithms in this domain.

Feature selection is inherently a combinatorial optimization problem (Kohavi and John, 1997). Conventional feature selection methods usually follow a greedy approach and choose top-ranking features on an individual level. This ignore the mutual dependency among the selected features. As a result of this, the optimal feature subset is sometimes difficult to find. The above mentioned five unsupervised feature selection algorithms except MCFS and UDFS follow the same methodology for obtaining the reduced feature set.

We attempt to incorporate the combinatorial effect, by adopting a graph theoretic approach utilising the notion of densest subgraph. The subgraph finding task is a known problem for a diverse number of applications like community mining, web mining, computational biology (Bahmani et al., 2012). Densest subgraph finding is a NP-hard problem. Recently, approximation algorithms for finding the densest subgraph have been devised in literature (Bahmani et al., 2012). Finding a subset of representative features by mining dense subgraph has also been addressed in Liu et al.

☆ This paper has been recommended for acceptance by S. Sarkar.

* Corresponding author. Tel.: +91 33 2575 3114; fax: +91 33 2578 3357.
*E-mail addresses:* sanghami@isical.ac.in (S. Bandyopadhyay), tapas.bhadra@isical.ac.in (T. Bhadra), pabitra@cse.iitkgp.ernet.in (P. Mitra), umaulik@cse.jdvu.ac.in (U. Maulik).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.12.008

(2011) and Mandal and Mukhopadhyay (2013). Liu et al. (2011) proposed a supervised method for obtaining the most informative features while Mandal and Mukhopadhyay (2013) used an unsupervised approach for obtaining the minimally redundant features. Here we have developed a new unsupervised feature selection technique based on the principle of densest subgraph finding followed by feature clustering.

We first obtain a graph representation by considering the entire feature set as the vertex set and having the inter-feature similarity as the corresponding edge weight. Here, the inter-feature similarity is computed using a normalized form of mutual information.

The densest subgraph finding approach has one major advantage that the vertices of this densest subgraph, i.e., the features of the reduced feature set, will be highly dissimilar. However, it is likely that these features may not be the optimal feature set. The reason behind this is that these features may not be the best representatives of the features that have been excluded, even though they are highly dissimilar to each other. To overcome this situation, a clustering approach is further applied on this densest subgraph for obtaining a better subgraph so that no important feature can be excluded from this set. The variance is used in the clustering phase to select the prototype feature while the same normalized mutual information is utilized for assigning each non-selected feature into its closest cluster representative. The subgraph thus obtained essentially contains a subset of the original features that can maximally represent the entire feature space. Thus our approach proceeds in a two-phase manner in which the first phase deals with finding out the densest subgraph while clustering the subgraph is performed in the second.

The remaining part of the paper is organized as follows: Section 2 discusses some preliminary concepts following which some of the existing unsupervised feature selection algorithms are discussed in Section 3. The proposed two-phase unsupervised feature selection algorithm is described in Section 4. Subsequently, the experiential design and the comparative results are provided in Section 5. Finally, some concluding comments are made in Section 6.

## 2. Preliminary concepts

This section describes some fundamental information and graph theory measures.

### 2.1. Density of a subgraph

Let $G = (V, E)$ be an unweighted undirected graph. The density of a subgraph $S \subseteq V$, denoted as $d(S)$, is defined as $d(S) = \frac{|E(S)|}{|S|}$, where $E(S)$ is the induced edge set of the subgraph S and $|S|$ is the cardinality of S.

The maximum density of the graph, denoted as $d^*(G)$, is defined as $d^*(G) = \max_{S \subseteq V}\{d(S)\}$. Similarly, the density of a subgraph $S \subseteq V$ within a weighted graph $G = (V, E)$ can also be defined as $d(S) = \frac{\sum_{e \in E(S)} w_e}{|S|}$, where $E(S)$ is the induced edge set of the subgraph $S$ and $w_e$ is the weight of the edge $e \in E(S)$.

### 2.2. Mutual information measures

#### 2.2.1. Entropy

Entropy of a random variable is the amount of uncertainty associated with it (Cover and Thomas, 2012). The entropy of a discrete variable $X$, denoted by $H(X)$, is defined as

$$H(X) = -\sum_{x \in X} p(x) \log_b p(x), \tag{1}$$

where $p(x)$ indicates the probability mass function of $X$. The value of $b$ is generally assumed to be 2.0 and this value is used in the present paper.

#### 2.2.2. Mutual information

Mutual information between two random variables measures how much information can be extracted through the knowledge of the other (Cover and Thomas, 2012). The value of mutual information becomes zero when the associated variables are completely independent whereas its higher value signifies their high mutual dependency. The mutual information between two discrete variables $X$ and $Y$, denoted as $I(X;Y)$, is defined as follows

$$I(X;Y) = \sum_{x \in X}\sum_{y \in Y} p(x,y) \log_b \left( \frac{p(x,y)}{p(x)p(y)} \right), \tag{2}$$

where $p(x), p(y)$ and $p(x,y)$ denote the probability mass function of $X$, the probability mass function of $Y$ and the joint probability mass function between $X$ and $Y$, respectively.

#### 2.2.3. Normalized mutual information

Mutual information has a disadvantage due to its non-comparability among variable pairs that have different mutual information values in various ranges. To overcome this, mutual information is often normalized into a closed interval, say $[0, 1]$.

Several researchers have used various methods to construct normalized mutual information. A few of them are mentioned below

$$\tilde{I}(X,Y) = \frac{2I(X;Y)}{H(X) + H(Y)}, \tag{3}$$

$$\hat{I}(X,Y) = \frac{I(X;Y)}{\min(H(X), H(Y))}, \tag{4}$$

$$I'(X,Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}. \tag{5}$$

Witten and Frank (2005) proposed the first one, known as symmetric uncertainty in the form of the weighted average of the two uncertainty coefficients. Strehl and Ghosh (2002) favoured the third form over the second one for ensembling several clusters due to the closeness to a normalized inner product in Hilbert space.

## 3. Review of unsupervised feature selection

Many of the earlier feature selection algorithms are based on supervised learning. Among the unsupervised feature selection approaches, data variance is the simplest measure for evaluating the discriminating power of a feature.

In the context of unsupervised feature selection algorithm, FSFS, proposed by Mitra et al. (2002), is a popular one. In this work, Mitra et al. (2002) proposed a new similarity measure, known as Maximal Information Compression Index (MICI) that was used to iteratively remove some number of features, say $k$, decrementing $k$ until no removal was possible. The MICI between two variables $x$ and $y$, denoted by $\lambda_2(x,y)$, was defined as follows

$$\lambda_2(x,y) = (var(x) + var(y)) - \sqrt{(var(x) + var(y))^2 - 4\,var(x)\,var(y)(1 - \rho(x,y)^2)}, \tag{6}$$

where $var(x)$, $var(y)$ and $\rho(x,y)$ denote the variance of $x$, the variance of $y$, and the correlation coefficient between $x$ and $y$, respectively.

A benefit of the approach is that it does not require any search which in turn makes the selection problem fast. However, this