ATLANTIS PRESS
Advances in Computer Science Research, volume 59
7th International Conference on Education, Management, Computer and Medicine (EMCM 2016)

A New Feature Selection Method Based on K-Nearest Neighbor Approach

Xianchang Wang[a*], Lishi Zhang[b] and Yonggang Ma[c]
School of Sciences, Dalian Ocean University, Dalian 116023, PR China
[a]wxcixll@sohu.com, [b]dwtag@sohu.com, [c]104774270@qq.com

Keywords: Feature selection; K-nearest neighbor; Unsupervised; Machine learning; Dimensionality reduction

Abstract. In many data analysis tasks, one is often confronted with very high dimensional data. Feature selection is an effective method to solve the problem with high dimensional data. The aim of feature selection is to reduce the number of features used in classification or recognition. This reduction is expected to improve the performance of classification and clustering algorithms in terms of speed, accuracy and simplicity. This paper proposes a new unsupervised feature selection algorithm which is based on K-nearest neighbor approach. The proposed algorithm evaluates the whole features according to the each sample in the dataset by the K-nearest neighbor approach. After that, the overall assessment is given based on the assessment of features for each sample. We evaluate the performance of the proposed unsupervised feature selection algorithm using the well-known UCI machine learning datasets , and the results illustrates the proposed algorithm is comparable with the traditional feature selection algorithm.

Introduction

Feature selection is a well-known pattern recognition problem, which is usually viewed as a data mining enhancement technique. This technique aims to reduce the number of features(variables) to be used, i.e., to reduce the entire feature space to a highly predictive subset of the space. This reduction may improve the performance of data mining algorithms to be used, in terms of speed, accuracy, and simplicity. In addition, because of this reduction, the identification of features which do not need to be stored, collected or bought, may bring financial savings[1].

The studies on feature selection can be classified two kinds: unsupervised feature selection method (such as the typical principle component analysis (PCA) method[2]) and supervised feature selection method (such as the classical linear discriminant analysis (LDA) method[3], neuro-fuzzy approaches based on an overall feature evaluation index[4]).

In the unsupervised learning, it is a nontrivial task to perform the feature selection in the absence of the ground-truth labels that could guide the assessment of the relevance and redundancy for each feature. Thus, the unsupervised feature selection problem becomes even more challenging than the supervised feature selection problem[5].

The studies on unsupervised feature selection can be further classified two kinds: finding the optimal feature subset and making some new features. Making some new features method will lose the original meaning of the dataset. Such as, PCA is concerned with summarizing the variance–covariance structure using a few linear combinations of the original set of variables (features)[2]. Thus, the study on unsupervised feature selection by finding the optimal feature subset is particularly important.

In the literature, there have been several representative methods that address the issue of the feature selection by finding the optimal feature subset. S. Tabakhi et al. presented an unsupervised feature selection method based on ant colony optimization (UFSACO), the method seeks to find the optimal feature subset through several iterations without using any learning algorithms[5]. Pabitra Mitra et al. described an unsupervised feature selection algorithm suitable for datasets, large in both dimension and size[6], the method is based on measuring similarity between features.

A fast feature selection algorithm for unsupervised massive datasets was proposed based on the incremental absolute reduction algorithm in traditional rough set theory in [7]. De Wang et al. proposed

a new feature selection framework to globally minimize the feature redundancy with maximizing the given feature ranking scores [8]. Clustering-guided sparse structural learning (CGSSL) is proposed by integrating cluster analysis and sparse structural analysis into a joint framework and experimentally evaluated for the unsupervised feature selection problem [9]. M. A. Ambusaidi et al. proposed an unsupervised feature selection algorithm, which is an enhancement over Laplacian score method [10].

However, the existing study on unsupervised feature selection by finding the optimal feature subset rank features by optimizing certain feature ranking criterions on the whole dataset. In this case, the selected features may be not suitable for some samples. Therefore, this paper proposes a new unsupervised feature selection by finding the optimal feature subset. Meanwhile, the proposed method can judge the selected features are propitious to which subset of data.

This paper is organized as follows: Section II proposes our unsupervised feature subset selection algorithm. Section III presents the experimental details and results. Finally, we conclude this paper by summarizing the work and future work in Section IV.

The Proposed Unsupervised Feature Subset Selection Algorithm

The data can be represented in the form $X=[x_{ij}]$ being an $n\times m$ matrix. Each column of X corresponds to a given feature (variable), and each row corresponds to a sample(pattern, data point). Let f_j ($j=1,...,m$) denote the j-th column (feature) of X , $x_i=[x_{i1},...x_{im}]$($i=1,...,n$) denote the i-th pattern, x_{ij} is the j-th feature value of x_i .

Definition 1: Let $N_{x_i}^k \subseteq X$ be the *k*-nearest neighbor of x_i , if $N_{x_i}^k$ satisfied that 1) $\forall x_p \in N_{x_i}^k$, $\forall x_q \in X \setminus N_{x_i}^k$, $d(x_p,x_i)\le d(x_q,x_i)$ and 2) $|N_{x_i}^k|=k$, where, the symbol $|\cdot|$ denotes the cardinality of a set, and $d(x_p,x_i)$ is the Euclidean distance.

Definition 2: Let $Nf_{jx_i}^k \subseteq X$ be the *k*-nearest neighbor of x_i on feature f_j , if $Nf_{jx_i}^k$ satisfied that 1) $\forall x_p \in Nf_{jx_i}^k$, $\forall x_q \in Nf_{jx_i}^k$, $d_{f_j}(x_p,x_i)\le d_{f_j}(x_q,x_i)$ and 2) $|Nf_{jx_i}^k|=k$, where, $d_{f_j}(x_p,x_i)=|x_{pj}-x_{ij}|$.

Definition 3: We define $E_{f_j}(x_i)=\left|N_{x_i}^k \cap Nf_{jx_i}^k\right|$ as the evaluation of feature f_j on sample x_i , the bigger value of $E_{f_j}(x_i)$ means the feature f_j is more important for sample x_i .

The proposed unsupervised feature subset selection algorithm is based on *k*-nearest neighbour of each sample, call KNNUFSSA, which includes the following four steps.

Step 1. For each sample x_i , calculating $E_{f_j}(x_i)$ for every feature f_j .

Step 2: For $j=1,...,m$, Let $W_{f_j}(x_i)$ be the ranking of f_j according to $E_{f_j}(x_i)$ in ascending order.

Step 3: Calculating $R_{f_j}=\sum_{i=1,...,n} W_{f_j}(x_i)$.

Step 4: Calculating $Score_{f_j}=R_{f_j}\Big/\sum_{k=1,...,m} R_{f_k}$ as the evaluation of feature f_j on the whole dataset.

Experimental Studies

In this section, we test the proposed method to the data coming from the UCI Repository of machine learning[11].The detail information of data is as follows.

1) Iris dataset: This is a well-known benchmark dataset which is widely used to test a learning algorithm in the field of machine learning. This dataset has 150 examples which are classified into three classes, i.e., Setosa, Versicolor and Virginical. Each example is characterized by four numerical features which are sepal length (SL), sepal width (SW), petal length (PL), and petal width (PW).

2) Pima diabetes dataset: The Pima Indian Diabetes dataset contains 768 examples. Each example representing a patient who may show signs of diabetes is described by eight features which are: a) number of times pregnant, b) plasma glucose concentration, c) diastolic blood pressure, d) triceps skin

fold thickness, e) two-hour serum insulin, f) body mass index, g) diabetes pedigree function, and h) age. There are 500 examples from patients who do not have diabetes and 268 examples from patients who are known to have diabetes.

3) Breast cancer diagnosis problem: The University of Wisconsin Breast Cancer data set consists of 699 patterns which are classified two classes, 458 benign examples and 241 malignant examples. Each example is described by nine features: a) clump thickness, b) uniformity of cell size, c) uniformity of cell shape, d) marginal adhesion, e) single epithelial cell size, f) bare nuclei, g) bland chromatin, h) normal nucleoli, and i) mitoses. In the dataset, the values of the sixth feature of 16 examples are missing. We neglect the 16 examples when conducting experiments.

In the whole experimental studies, the class labels are neglected.

For the iris dataset, given the parameter $k=50$, thus, $W_{f_1}(x_1)=2$, $W_{f_2}(x_1)=1$, $W_{f_3}(x_1)=4$, $W_{f_4}(x_1)=3$, $W_{f_1}(x_{150})=2$, $W_{f_2}(x_{150})=4$, $W_{f_3}(x_{150})=1$, $W_{f_4}(x_{150})=3$. $R_{f_1}=\sum_{i=1,\ldots,150}W_{f_1}(x_i)=304$, $R_{f_2}=\sum_{i=1,\ldots,150}W_{f_2}(x_i)=251$, $R_{f_3}=\sum_{i=1,\ldots,150}W_{f_3}(x_i)=465$, $R_{f_4}=\sum_{i=1,\ldots,150}W_{f_4}(x_i)=480$.

Due to $Score_{f_1}=0.20$, $Score_{f_2}=0.17$, $Score_{f_3}=0.31$, $Score_{f_4}=0.32$, thus, the ranking of features as follows: f_4, f_3, f_1, f_2.

For the pima dataset, given the parameter $k=250$, thus, $W_{f_1}(x_1)=1$, $W_{f_2}(x_1)=8$, $W_{f_3}(x_1)=5$, $W_{f_4}(x_1)=7$, $W_{f_5}(x_1)=3$, $W_{f_6}(x_1)=6$, $W_{f_7}(x_1)=4$, $W_{f_8}(x_1)=2$, $W_{f_1}(x_{768})=1$, $W_{f_2}(x_{768})=7$, $W_{f_3}(x_{768})=8$, $W_{f_4}(x_{768})=5$, $W_{f_5}(x_{768})=6$, $W_{f_6}(x_{768})=3$, $W_{f_7}(x_{768})=2$, $W_{f_8}(x_{768})=4$. $R_{f_1}=3390$, $R_{f_2}=3736$, $R_{f_3}=3882$, $R_{f_4}=3664$, $R_{f_5}=3371$, $R_{f_6}=3243$, $R_{f_7}=3108$, $R_{f_8}=3254$.

Due to $Score_{f_1}=0.123$, $Score_{f_2}=0.135$, $Score_{f_3}=0.14$, $Score_{f_4}=0.133$, $Score_{f_5}=0.122$, $Score_{f_6}=0.117$, $Score_{f_7}=0.112$, $Score_{f_8}=0.118$, thus, the ranking of features as follows: f_3, f_2, f_4, f_1, f_5, f_8, f_6, f_7.

For the Breast cancer diagnosis data, given the parameter $k=200$, thus, $W_{f_1}(x_1)=9$, $W_{f_2}(x_1)=5$, $W_{f_3}(x_1)=8$, $W_{f_4}(x_1)=7$, $W_{f_5}(x_1)=4$, $W_{f_6}(x_1)=3$, $W_{f_7}(x_1)=6$, $W_{f_8}(x_1)=2$, $W_{f_9}(x_1)=1$, $W_{f_1}(x_{683})=8$, $W_{f_2}(x_{683})=3$, $W_{f_3}(x_{683})=7$, $W_{f_4}(x_{683})=6$, $W_{f_5}(x_{683})=5$, $W_{f_6}(x_{683})=2$, $W_{f_7}(x_{683})=4$, $W_{f_8}(x_{683})=9$, $W_{f_9}(x_{683})=1$. $R_{f_1}=5287$, $R_{f_2}=3657$, $R_{f_3}=3700$, $R_{f_4}=3606$, $R_{f_5}=3709$, $R_{f_6}=3439$, $R_{f_7}=2913$, $R_{f_8}=2291$, $R_{f_9}=2133$.

Due to $Score_{f_1}=0.172$, $Score_{f_2}=0.119$, $Score_{f_3}=0.120$, $Score_{f_4}=0.117$, $Score_{f_5}=0.121$, $Score_{f_6}=0.112$, $Score_{f_7}=0.095$, $Score_{f_8}=0.075$, $Score_{f_9}=0.069$, thus, the ranking of features as follows: f_1, f_5, f_3, f_2, f_4, f_6, f_7, f_8, f_9.

We compare our proposed method KNNUFSSA with OFEI[12], FQI[12], OFFSS[13], and MIFS[13], see the Table 1. The result illustrates KNNUFSSA is comparable with the other feature subset selection method, although OFEI, FQI, OFFSS, and MIFS are the supervised feature subset selection learning method.

Summary

This paper proposes a new unsupervised feature selection method, meanwhile, the proposed method can judge the selected features are propitious to which subset of data. The experimental results illustrate KNNUFSSA is comparable with the other feature subset selection method.

Table 1 Feature subset selection by different approaches.

	iris	pima	breast cancer
OFFSS	$\{4,3\}$	$\{2,6,7\}$	$\{6,3,1,2\}$
MIS	$\{4,3\}$	$\{2,6,8\}$	$\{6,3,2,7\}$
FQI	$\{4,3\}$	$\{8,2,1\}$	$\{6,1,8,3\}$
OFEI	$\{4,3\}$	$\{2,3,6\}$	$\{6,1,3,2\}$
KNNUFSSA	$\{4,3\}$	$\{3,2,4,1\}$	$\{1,5,3,2\}$

Acknowledgements

This work was supported in part by the Scientific Research Program for The Education Department of Liaoning Province of China (Grant No. L201606, L201615) and Doctor Project (Grant No. 38/103816003).

References

[1] C. W. D. Justin, R. J. Victor, Feature subset selection with a simulated annealing data mining algorithm, J. Intell. Inform. Syst. 9 (1997) 57-81.

[2] K. Fukunaga, Introduction to Statistical Pattern Recognition, New York: Academic, 1972.

[3] I. T. Joliffe, Principal Component Analysis, New York: Springer-Verlag, 1986.

[4] R. R. Yager, L. A. Zadeh, Fuzzy Sets, Neural Networks, and Soft computing, New York: Van Nostrand-Reinhold, 1994.

[5] S. Tabakhi, P. Moradi, F. Akhlaghian, An unsupervised feature selection algorithm based on ant colony optimization, Engineering Applications of Artificial Intelligence, 32(2014) 112-123.

[6] P. Mitra, C. A. Murthy, S. K. Pal, Unsupervised feature selection using feature similarity, IEEE Transactions on Pattern Analysis & Machine Intelligence 24 (2002) 301-312.

[7] H. Bai, J. Wang, L. I. Deyu, et al., Fast unsupervised feature selection algorithm based on rough set theory, Journal of Computer Applications, 35 (2015) 2355-2359.

[8] D. Wang, F. Nie, H. Huang, Feature selection via global redundancy minimization, IEEE Transactions on Knowledge & Data Engineering, 27(2015) 2743-2755.

[9] Z. Li, J. Liu, Y. Yang, X. Zhou, H. Lu, Clustering-guided sparse structural learning for unsupervised feature selection, IEEE Transactions on Knowledge & data Engineering, 26 (2014) 2138 - 2150.

[10] M. A. Ambusaidi, X. He, P. Nanda, Unsupervised feature selection method for intrusion detection system, IEEE Trustcom/bigdatase/ispa. IEEE Computer Society, (2015) 2497-2507.

[11] A. Frank, A. Asuncion, UCI machine learning repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science 213 (2010).

[12] R. K. De, N. R. Pal, S. K. Pal, Feature analysis: Neural network and fuzzy set theoretic approaches, Pattern Recognition, 30 (1997) 1579–1579.

[13] E. C. C. Tsang, D. S. Yeung, X. Z. Wang, OFFSS: optimal fuzzy-valued feature subset selection, IEEE Transactions on Fuzzy Systems, 11 (2003) 202-213.