IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10137

Supervised, Unsupervised and Semi-supervised learning

Sujata Gawade[1], Pournima Kamle[2]

Lecturer, Department of Computer Technology , BVIT, Navi Mumbai, India [1]

Lecturer, Department of Computer Technology , BVIT, Navi Mumbai, India [2]

Abstract: Feature selection is a big task and its challenge for high dimensional data. Semi-supervised feature selection is a combination of supervised and unsupervised data. Supervised data means labelled data and unsupervised data means unlabelled data. In semi-supervised feature selection unlabelled data is more than labelled data. Supervised learning means well labelled data that means data is already labelled with correct answer. Unsupervised data means that data is neither labelled nor classified and allowing algorithm to process without guidance. Supervised learning use processes like regression and classification. Algorithms used for unsupervised learning clustering and association. In supervised learning optimized performance criteria with the help of previous experience.

Keywords: Supervised, Semi-supervised, labelled, and unlabelled.

I. INTRODUCTION

In high dimensional data, dimensionality reduction is an important task. It can be also used to reduce the dimensionality of the real data and optimize performance. To eliminate irrelevant as well as redundant features, or by efficiently merging real features to produce a minor set of them with added discriminative power [1]. For dimensionality reduction there are two methods: Feature selection and Feature extraction. Feature extraction method in which merging of real features and due to which dimensionality reduction is possible. For labeled data accessibility, feature extraction method can be classified as managed and verified. Fisher Linear Discriminant [2] FLD is an example of supervised learning and it is based on class labels. In unsupervised feature extraction method global covariance structure method is used when class labels are not present. The Principal Component Analysis method is an example of unsupervised learning [3].Feature selection means to identify relevant features in supervised and unsupervised data. In unsupervised data class labels are not present for feature relevance; it can be done with the help of separation or variance. In Semi-supervised feature selection supervised data as well as unsupervised data. It is called small-labeled sample problem means in which supervised data is smaller in quantity and unlabeled data is more in size. In this case supervised feature selection algorithm didn't used for feature selection. It is not possible to select irrelevant features and to remove redundant features; in this case unsupervised data is larger in size so that semi-supervised feature selection is
required. The gap between unsupervised and supervised feature selection seems tough to close as one works with absence of class labels and other with the class labels. If we change the viewpoint and put least focus on class labels, both unsupervised and supervised feature selection can be viewed as an effort to select features that are consistent with the target concept. In supervised learning the main concept is related to class labels, while in unsupervised learning the main concept is related to the structures of the data. In supervised and unsupervised learning, the main motto is basically to divide instances into various subsets according to different features. The challenge now is how to develop a unified depiction based on which different types of features can be measured. For feature selection pairwise constraints are selected. Pairwise constraint selection is used for both supervised as well as unsupervised learning. It is used to describe relationship among instances. Basically feature relevance can be expressed in different forms likewise pairwise constraints or class labels, or other prior information. There are pairs of instances which are belonging to the same class known as (must-link constraint) and different classes (cannot link constraint). We focus on domain knowledge in the form of pairwise constraints, i.e. pairs of instances known as belonging to the same class (must-link constraints) or different classes (cannot-link constraints). For image retrieval constraints arise naturally. Pairwise constraints is practical than class labels because pairwise constraints is derived from labeled data. Pairwise constraints derived automatically without human interference.

II. MOTIVATION

Semi-supervised feature selection is suffered from the problem of presence unlabeled and labeled instances together. Unlabeled and labeled cases are used to represent to solve objective concept. Semi-supervised feature selection required to use both partitions of data

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

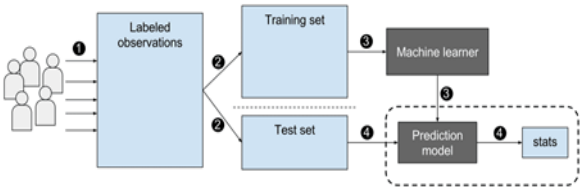
International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10137

III.ARCHITECTURE

Supervised learning: Supervised learning is a type of machine learning task that records an input to an output based on example input-output pairs. There is labeled training data.



It concludes a function from labeled data means that data is already trained. Training data means pair of an input object and a desired output value. A supervised learning algorithm investigates training data and produces secondary function which can be used for mapping new examples. Class labels are mainly used for instances. This requires learning algorithm to simplify training data in unseen situations. A classification means it is just like category when you consider about colour such as "purple" or "violet" or "pass" and "fail".

Regression: A regression problem is when the output variable is a real value, such as "dollars" or "weight".

Regression technique predicts single output value using training data.

For example you can use regression to predict old car price according to average, colour, total mileage etc.

Neural network, support vector machine, linear and logistic regression, random forest and classification trees are used.

Unsupervised learning:

Unsupervised machine learning is a process where not requirement of supervised model. It deals with unlabeled data. In unsupervised learning no. of kind of unknown patterns in data. Unsupervised methods are used for finding features which are helpful for categorization.

Unsupervised learning classified into two types:

1 .Clustering

2. Association

Clustering:

Clustering is the most important concept when it comes to unsupervised learning. It mostly deals with finding a pattern or structure in a collection of unclassified data. These clustering algorithms will process data and find natural clusters or groups if they exist in the data. Clustering is used to adjust the granularity of groups.

There are different types of clustering:

Exclusive (partitioning):

Data are grouped in such a way that it belongs to one group only.

Agglomerative

The iterative unions between the two nearest clusters reduce the number of clusters. Every data is a cluster.

Example: Hierarchical clustering.

Overlapping

In this technique, fuzzy logic is used for cluster. There are separate degrees of membership for each point while making cluster. Data is associated with appropriate membership value.

Example: Fuzzy C-Means

Probabilistic:

Probability distribution uses to create the clusters in this technique.

Example:

Man's shoe

Women's shoe

Clustering types:

1. Hierarchical clustering
2. K-NN clustering
3. K-means clustering
4. Principal Component Analysis

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10137

Hierarchical clustering:

In this technique there is hierarchy of clusters .There is grouping of two same clusters. Two close clusters form only one cluster.

K-means Clustering

K means it is an iterative clustering algorithm which used to find the highest value for every iteration. There is specific number of clusters are selected. In this clustering method, cluster the data points into k groups. A larger k means smaller groups with more granularity in the same way. A lower k means larger groups with less granularity.

Principal Component Analysis:

It is non-parametric stastical technique used for dimensionality reduction. High dimensionality has a large number of features ,main problem with high dimension is model overfitting. PCA can be used for image compression and filter noisy datasets.

Association:

Association rules used to establish relations amongst data objects inside large databases. This unsupervised technique is about discovering stimulating relationships between variables in large databases. For example, people that buy a new car most likely to buy new accessories.

Semi-supervised learning:

In high dimensional data there is presence of large amount of data and in which some of data is labeled is called semi-supervised feature selection.

Example of semi-supervised feature selection is photo archive and there are images of dog ,cat and person some of images are labeled and majority are unlabeled.

IV.CONCLUSION

Semi-supervised feature selection is large amount of unlabeled data and small amount of labeled data.

REFERENCES

[1] Z. Zhao and H. Liu, Spectral Feature Selection for Data Mining (DataMining and Knowledge Discovery Series). Boca Raton, FL,USA:Chapman and Hall-CRC, 2012

[2] R. Fisher, "The use of multiple measurements in taxonomic problems," Ann. Eugen, vol. 7, no. 2, pp. 179–188, Sept. 1936.

[3] I. Jolliffe, Principal Component Analysis. New York, NY, USA:Springer, 2002.[4] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reductionby local linear embedding," Science, vol. 290, no. 5500,pp. 2323–2326, Dec. 2000.

[5] X. He and P. Niyogi, "Locality preserving projections," in Proc.NIPS, 2004.[6] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectraltechniques for embedding and clustering," in Proc. NIPS, 2002.

[8] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," J. Mach. Learn. Res., vol. 5, pp. 845–889, Aug. 2004.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification .New York, NY, USA: Wiley Interscience, 2000.

[10] M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of relief and relieff," Mach. Learn., vol. 53, no. 1–2, pp. 23–69, 2003.

[11] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," J. Mach. Learn. Res., vol. 5, pp. 1205–1224, Oct. 2004.

[12] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in Proc. 24th Int. Conf. Mach. Learn., Corvallis, OR, USA, 2007.

[13] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt,"Feature selection via dependence maximization," J. Mach. Learn. Res., vol. 13, no. 1, pp. 1393–1434, Jan. 2012.

[14] Z. Zhao and H. Liu, "Semi-supervised feature selection via spectral analysis," in Proc. SIAM Int. Conf. Data Mining, Tempe,AZ, USA, 2007, pp. 641–646.

[15] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a Mahalanobis metric from equivalence constraints," J. Mach. Learn.Res., vol. 6, pp. 937–965, Jan. 2005.

[17] D. Zhang, S. Chen, and Z. Zhou, "Constraint score: A new filtermethod for feature selection with pairwise constraints," Pattern Recognit., vol. 41, no. 5, pp. 1440–1451, 2008.[18] K. Benabdeslem and M. Hindawi, "Constrained Laplacian score for semi-supervised feature selection," in Proc. ECML-PKDD,Athens, Greece, 2011, pp. 204–218.

[19] M. Hindawi, K. Allab, and K. Benabdeslem, "Constraintselection based semi-supervised feature selection," in Proc. IEEE ICDM, Vancouver, BC, Canada, 2011, pp. 1080–1085.

[20] Sam T. Roweis and Lawrence K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," Science 290, 2323 (2000).

[21] Huan Liu, "Feature Selection: An Ever Evolving Frontier in Data Mining," JMLR: Workshop and Conference Proceedings 10

[22] Mohammed Hindawi, Kaïs Allab, Khalid Benabdeslem,"Constraint Selection based Semi-supervised Feature Selection".