Feature Selection and Extraction in Data mining

Aparna.U.R
PG Scholar
Dept.of Computer Science Engineering
Jyothi Engineering College
Kerala,India

Shaiju Paul
Dept. Of Computer Science Engineering
Jyothi Engineering College
Kerala,India

***Abstract*—Data mining is the process of extraction of relevant information from a collection of data. Mining of a particular information related to a concept is done on the basis of the feature of the data. The accessing of these features hence for data retrieval can be termed as the feature extraction mechanism. Different type of feature extraction methods are being used. The feature selection algorithm should fit with the offline as well as on-line mining .**

***Index Terms*—Feature extraction, feature selection, data mining.**

I. INTRODUCTION

Feature selection can be termed as the process of selecting a particular feature from a huge collection of features. All the features are residing within the data. So we need to select a particular feature from huge set of features that are residing with in the dataset .Feature selection plays an important role in the machine learning and data mining. Machine learning is a subfield of computer science that evolved from the study of the pattern recognition and computational learning theory in artificial intelligence. In machine learning, feature selection is also termed as variable selection or attribute selection. Feature selection technique are used for the following reasons:

a)Simplification of the data models to make easier interpretation.

b)Reduction in the training time.

c)Reducing over-fitting.

II. MOTIVATION

The real world offers many challenging problems. The extraction of information is being done from an enormous number of database. Data base are huge in size. Databases are also growing at an unprecedented rate. Decision need to be done rapidly and the decisions must be made with maximum knowledge. This leads to the complexity of time and space. A proper feature extraction mechanism is required to properly extract the information.

III. DATA MINING

Data mining deals with the extraction of information from huge collection of data. Data mining deals with extremely large databases. Data mining is a combination of several approaches. Data mining deals with machine learning , data base, visualization, applied statistics, pattern recognition, parallel algorithms, high performance etc. Knowledge discovery in data base leads with data warehousing, data selection, data preprocessing, data transformation, data mining, Interpretation and evaluation.

The amount of data being created as well as utilized is increasing very rapidly. The information should be extracted from a huge collection of data. The dimensionality of the data is increasing in such a manner that it is difficult for extracting a particular information rapidly with the currently available methods. Extraction of a particular information should be done from a huge collection of database. The size of the database are huge and they are also increasing rapidly at an unprecedented rate.

As the amount of data in the databases rises, data that is required for a proper analysis has also increased. The increase in the dimensionality of the data leads to complexity in the data mining task. As the dimensionality of the data increases, the computational cost required for analysis also grows rapidly. Hence, a better mechanism is required to deal with high dimensional data. Reduction in the number of features can be done by proper Feature Selection and Extraction mechanism.

IV. FEATURE SELECTION

The extraction of information can be done easily by feature selection mechanism. Hence the feature selection mechanism plays an important role in data mining. For performing a data mining operation we are dealing with redundant feature. These features may be irrelevant . This can lead to the increase in complexity. Hence reducing the number of irrelevant/redundant features plays an important role. Feature selection mechanism is done so as to eliminate the redundant features and extract the information. Feature should be selected in such a way that the:

i)The accuracy and performance should not be affected.

ii)output must be the same .

A better feature selection mechanism helps in facilitating the data visualization, data understanding, reducing the storage

978-1-5090-4556-3/16/$31.00 ©2016 IEEE

requirements and utilization in time and in reducing the dimensionality.

Feature selection mechanism deals with two main steps:

-Feature generation and

-Feature evaluation

Feature subset generation is the process of selection of subset of features from huge collection. Feature evaluation deals with the evaluation of the subset of features in such a way so as to fit the requirements. This evaluation can be done on the basis of dependent and independent measures. Dependence measure evaluates the feature subset by monitoring the performance of algorithm applied on it. Evaluation without the application of any learning algorithm is done in independent measures. The different types of independent measures are:

Distance Measures: Distance measures deals with the measuring of similarity or distance between two points. These points having smaller distance between them have similar features.

Information Measures: Deals with the measurement of information that a selected feature provides. An infrequently occurring event provides more information than frequently occurring event.

Dependency Measure: Dependency measure deals with measuring the dependence between two random variables.

Consistency Measure: The concept of consistency measures was introduced to evaluate the distance of a given feature from consistency state.

Feature selection works by removing features that are not relevant or are redundant. Feature selection algorithms can be classified into three categories. They are filters, wrappers and embedded techniques.

Filters work by extracting features from the data without any learning mechanism. Wrappers works by application of learning techniques. The embedded techniques deal with a combination of approaches. It combines the feature selection step and classifier construction.

A. Filters

Filters work without taking classifier into consideration. Filters can be classified into multi-variate and uni-variate. Multi-variate methods are able to find the relationship among features. Uni-variate methods consider each features separately.

B. Wrappers

Wrappers considers model hypothesis into account by training and testing in the feature space. Wrappers perform better in selecting features. The dependencies of the features are considered by the wrapper method.

C. Embedded Methods

Embedded methods integrate the feature selection process into the model training process. They are usually faster than the wrapper methods and able to provide suitable feature subset for the learning algorithm.

V. FEATURE EXTRACTION

Feature extraction creates new variables as a combination of others to reduce the dimensionality of the selected features. Feature extraction can be classified into two:

a)Linear and

b)Non-linear

Feature extraction has been one of the most important issues of pattern recognition. Most of the feature extraction literature has centered on finding linear transformations, which map the original high-dimensional sample space into a lower-dimensional space that hopefully contains all discriminatory information. The principal motivation behind dimensionality reduction by feature extraction is that it may reduce the worst effects of the curse of dimensionality. Also linear feature extractions techniques are often used as pre-processors before more complex nonlinear classifiers.

VI. LEARNING APPROACHES

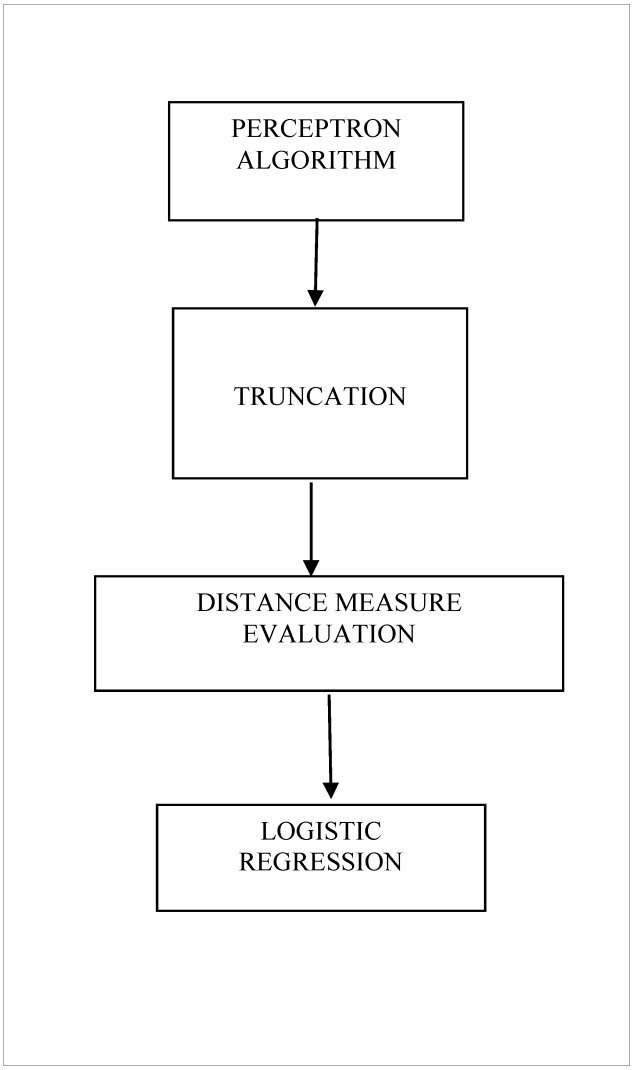
Learning mechanisms can be classified into Supervised, Unsupervised & Reinforced learning. Supervised Learning deals with the learning of classification system that we already created. Unsupervised deals with learning something but the procedure of doing it is not specified. In reinforcement learning, the learner is not told about which action to be taken, but the learner by itself needs to discover actions to yield the outcome.

VII.PROPOSED WORK

The proposed work is based on supervised learning. Perceptron is an algorithm for supervised learning of binary classifiers. It helps to decide whether a feature belongs to one class or another. It is a type of linear classifier, i.e. a classification algorithm that makes its predictions based on a linear predictor function combining a set of weights with the feature vector. Perceptron algorithm is an online learning method. It is a supervised learning method. The proposed work is framed to be with the perceptron algorithm. Based on the input requirements, a weight is set. Those data which are above the weight prescribed get excited and moves to the next node. Those data are selected and the rest are rejected.

Usually most data mining takes in such a way that the data retrieval will takes place with some features that are similar. Most of the present features which are used for extraction are linked to the previous features in some aspect. There is a familiarity in the data extracted. So, the features remain closer. These selected features are truncated to the nearest possible value. So, all the features that remains closer have a common truncated value.

After the truncation, distance measure evaluation is done. Those features which are closer can have little difference among the distance calculated. Logistic regression on this features hence can improve the feature selection mechanism. Hence the feature extraction can be done efficiently.



VII. REFERENCES

[1] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, *"Distributional Word Clusters versus Words for Text Categorization,"* J. Machine Learning Research, vol. 3, pp. 1183-1208, 2003.

[2] J. Bi, K.P. Bennett, M.J. Embrechts, C.M. Breneman, and M. Song, "*Dimensionality Reduction via Sparse Support Vector Machines,"* J. Machine Learning Research, vol. 3, pp. 1229-1243, 2003.

[3] G. Cavallanti, N. Cesa-Bianchi, and C. Gentile, *"Tracking the Best Hyperplane with a Simple Budget Perceptron,"* Machine Learning, vol. 69, nos. 2-3, pp. 143-167, 2007.

[4] N. Cesa-Bianchi, S. Shalev-Shwartz, and O. Shamir, *"Efficient Learning with Partially Observed Attributes,"* J. Machine Learning Research, vol. 12, pp. 2857-2878, 2011.

[5] A.B. Chan, N. Vasconcelos, and G.R.G. Lanckriet, "*Direct Convex Relaxations of Sparse SVM,"* Proc. 24th Int'l Conf. Machine Learning (ICML '07), pp. 145-153, 2007.

[6] K. Crammer, O. Dekel, J. Keshet, S. ShalevShwartz, and Y. Singer, *"Online Passive-Aggressive Algorithms,"* J. Machine Learning Research, vol. 7, pp. 551-585, 2006.

[7] K. Crammer, M. Dredze, and F. Pereira, *"Exact Convex Confidence-Weighted Learning,"* Proc. Advances in Neural Information Processing Systems (NIPS '08), pp. 345-352, 2008.

[8] K. Crammer, A. Kulesza, and M. Dredze, *"Adaptive Regularization of Weight Vectors,"* Proc. Advances in Neural Information Processing Systems (NIPS '09), pp. 414-422, 2009.

[9] M. Dash and V. Gopalkrishnan, *"Distance Based Feature Selection for Clustering Microarray Data,"* Proc. 13th Int'l Conf. Database Systems for Advanced Applications (DASFAA '08), pp. 512-519, 2008.

[10] M. Dash and H. Liu, *"Feature Selection for Classification," Intelligent Data Analysis*, vol. 1, nos. 14, pp. 131-156, 1997.

[11] X. Wu, K. Yu, H. Wang, and W. Ding, *"Online Streaming Feature Selection,"* Proc. Int'l Conf. Machine Learning (ICML '10), pp. 11591166, 2010.

[12] Z. Xu, R. Jin, J. Ye, M.R. Lyu, and I. King, *"NonMonotonic Feature Selection,"* Proc. Int'l Conf. Machine Learning (ICML '09), p. 144, 2009.

[13] Z. Xu, I. King, M.R. Lyu, and R. Jin, *"Discriminative Semi-Supervised Feature Selection via Manifold Regularization*," IEEE Trans. Neural Networks, vol. 21, no. 7, pp. 1033-1047, July 2010.

[14] S. Yang, L. Yuan, Y.-C. Lai, X. Shen, P. Wonka, and J. Ye, *"Feature Grouping and Selection over an Undirected Graph,"* Proc. 18th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '12), pp. 922-930, 2012.

[15] Y. Yang, H.T. Shen, Z. Ma, Z. Huang, and X. Zhou, *"l2;1-Norm Regularized Discriminative Feature Selection for Unsupervised Learning,"* Proc. 22nd Int'l Joint Conf. Artificial Intelligence (IJCAI '11), pp. 15891594, 2011.

[16] L. Yu and H. Liu, "*Feature Selection for HighDimensional Data: A Fast Correlation-Based Filter Solution*," Proc. Int'l Conf. Machine Learning (ICML '03), pp. 856-863, 2003.

[17] P. Zhao and S.C.H. Hoi, *"OTL: A Framework of Online Transfer Learning,"* Proc. Int'l Conf. Machine Learning (ICML '10), pp. 12311238, 2010.

[18] P. Zhao and S.C.H. Hoi, *"Bduol: Double Updating Online Learning on a Fixed Budget,"* Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML/PKDD '12), no. 1, pp. 810-826, 2012.

[19] P. Zhao, S.C.H. Hoi, and R. Jin, *"Double Updating Online Learning,"* J. Machine Learning Research, vol. 12, pp. 1587-1615, 2011.

[20] P. Zhao, S.C.H. Hoi, R. Jin, and T. Yang, *"Online Auc Maximization,"* Proc. Int'l Conf. Machine Learning (ICML '11), pp. 233-240, 2011.