Vol.2 Issue.5, Pg.: 102-105
May 2014

INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

FEATURE SELECTION FOR HIGH-DIMENSIONAL DATA BY USING KRUSKALS ALGORITHM

Ms.R.Devipriya (M.E), Department of CSE

Muthayammal Engineering College, India

devipriyaatrs@gmail.com

Abstract

Feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy and improving results comprehensibility. This process improved by cluster based FAST Algorithm using MST construction. The instances that define a neighbourhood are used as aggregation points to capture feature relevance. Irrelevant feature subspaces within the neighbourhood are used as evidences of negative relevance. of FAST has a high probability of producing a subset of useful and independent features. The proposed algorithm not only reduces the number of features, but also improves the performances of the four well-known different types of classifiers such as the probability-based Naive Bayes, the tree-based C4.5, the instance-based IB1, and the rule-based RIPPER before and after feature selection SCRAP was compared with the RELIEF filtering scheme and was found to be the better scheme on all the different classes of learners. an Instance Based filter approach to Feature Selection called Selection Construction Ranking using Attribute Pattern (SCRAP). This work captures the benefit of feature selection for Instance Based Learners like Nearest Neighbors. We further analyze the generality of the feature sub sets produced by Instance Based filtering by running them on 3 classes of learners.

Keywords: Feature Subset Selection, RELIEF filtering, Naïve Bayes

1. Introduction

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. A side from the raw analysis step, it involves database and data management in aspects, model and inference considerations, metrics, complexity considerations, ,post-processing.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system.

1.1 DATA CLUSTERING

Clustering is useful in several machine learning and data mining tasks including: image segmentation, information retrieval, pattern recognition, pattern classification, network analysis, and so on.

It can be seen as either an exploratory task or pre-processing step. If the goal is to explore and reveal the hidden patterns in the data, clustering becomes a stand-alone exploratory task by itself. However, if the generated clusters are going to be used to facilitate another data mining or machine learning task, clustering will be a pre-processing step in this case. There are many clustering methods in the literature. These methods can be categorized broadly into: partitioning methods, hierarchical methods, and density-based methods.

These methods are sometimes divided into *partitioning* methods, in which the classes are mutually exclusive, and the less common *clumping* method, in which overlap is allowed. Each object is a member of the cluster with which it is most similar, however the threshold of similarity has to be defined. The hierarchical methods produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains. The hierarchical methods can be further divided into *agglomerative* or *divisive* methods.

1.2 PARTITIONING METHODS

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases, e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids

1.3 PREPROCESSING

Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data preprocessing that includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set.

1.4 MST CONSTRUCTION

A minimum spanning tree (MST) or minimum weight spanning tree is then a spanning tree with weight less than or equal to the weight of every other spanning tree. More generally, any undirected graph (not necessarily connected) has a minimum spanning forest, which is a union of minimum spanning trees for its connected components. Finding the smallest edge can be done at the same time as updating minimum spanning tree.

When building a minimum spanning tree on a complete graph , an algorithm which has a complexity based on the number of edges must have a complexity better than O(M) to beat Prim's algorithm.

1.5 TREE PARTITION

Each tree in the MST represents a cluster. In this module, we apply graph theoretic clustering methods to features.

In particular, we adopt the minimum spanning tree (MST) based clustering algorithms, because they do not assume that data points are grouped around centres or separated by a regular geometric curve and have been widely used in practice.

1.6 FEATURE SELECTION

Feature subset selection was split up into two parts, subset searching and criterion functions. For both parts, the common algorithms were introduced and analyzed. Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features.

APPLICATIONS

Data clustering has immense number of applications in every field of life. One has to cluster a lot of thing on the basis of similarity either consciously or unconsciously. So the history of data clustering is old as the history of mankind.

In computer field also, use of data clustering has its own value. Especially in the field of information retrieval data clustering plays an important role. Some of the applications are listed below.

PROPOSED WORK

The feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features is proposed. This is because irrelevant features do not contribute to the predictive accuracy and redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s) and construct minimum spanning trees to evaluate whether two sets of n-dimensional data are from the same distribution.

The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature elimination A minimum spanning tree is built across the data points, and edges which connect data from one distribution to the other are removed. If many edges are removed, then the data from the distributions are mixed up together, and so they must come from the same distribution. A minimum-spanning tree is a sub-graph of a weighted, connected and undirected graph. It is acyclic, connects all the nodes in the graph, and the sum of all of the weight of all of its edges is minimum. That is, there is no other spanning tree, or sub-graph which connects all the nodes and has a smaller sum.

Vol.2 Issue.5, Pg.: 102-105
May 2014

CONCLUSION

Feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy. In this cluster based feature subset selection algorithm is used to select the features in efficient and accuracy. The cluster based feature subset selection algorithm involves (i) removing irrelevant features, (ii) constructing a minimum spanning tree from relative ones, (iii) partitioning the MST and selecting representative features. In the proposed system FAST algorithm is used. A cluster consists of features.

Each cluster is treated as a single feature and thus dimensionality is drastically reduced.

The KRUSKALS algorithm obtained the best proportion of selected features.

The best runtime and the best classification accuracy for Naïve Bayes, C4.5, and RIPPER, and the second best classification accuracy for IB1. The clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features.

REFERENCES

[1] Almuallim H. and Dietterich T.G., Algorithms for Identifying Relevant Features, In Proceedings of the 9th Canadian Conference on AI, pp 38-45, 1992.

[2] Almuallim H. and Dietterich T.G., Learning boolean concepts in the presence of many irrelevant features, Artificial Intelligence, 69(1-2), pp 279- 305, 1994.

[3] Arauzo-Azofra A., Benitez J.M. and Castro J.L., A feature set measure based on relief, In Proceedings of the fifth international conference on Recent Advances in Soft Computing, pp 104-109, 2004.

[4] Baker L.D. and McCallum A.K., Distributional clustering of words for text classification, In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, pp 96-103, 1998.

[5] Battiti R., Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks, 5(4), pp 537- 550, 1994.

[6] Bell D.A. and Wang, H., A formalism for relevance and its application in feature subset selection, Machine Learning, 41(2), pp 175-195, 2000.

[7] Biesiada J. and Duch W., Features election for high-dimensionaldatała Pearson redundancy based filter, AdvancesinSoftComputing, 45, pp 242C249, 2008.

[8] Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., On Feature Selection through Clustering, In Proceedings of the Fifth IEEE international Conference on Data Mining, pp 581-584, 2005.

[9] Cardie, C., Using decision trees to improve case-based learning, In Proceedings of Tenth International Conference on Machine Learning, pp 25-32, 1993.

[10] Chanda P., Cho Y., Zhang A. and Ramanathan M., Mining of Attribute Interactions Using Information Theoretic Metrics, In Proceedings of IEEE international Conference on Data Mining Workshops, pp 350-355.