Asian Journal of Information Technology 15 (19): 3730-3733, 2016
ISSN: 1682-3915

An Efficient Feature Subset Algorithm for High Dimensional Data

T. Divya and B. Vijaya Babu
Department of CSE, KL University, Green fields, Vaddeswaram,
Guntur, 522502 Andhra Pradesh, India

Abstract: Selection of feature subset is an efficacious way for dimensionality contraction, elimination of inappropriate data ascending learning accurateness and improving result unambiguousness. Numerous feature subset selection design have been planned and studied for machine learning applications. Feature subgroup selection can be analyzed as the process of admit and eliminating as many improper and redundant features as bright since inappropriate features do not put in to the anticipating accurateness and superfluous characteristics do not react to getting an enhanced predictor for that they make accessible mainly instruction which is by now present in earlier feature. We build up a novel design that can capably and effortlessly deal with both incorrect and superfluous characteristics and get hold of a superior character subset. Based on the minimum spanning tree method, we confirm a FAST algorithm. The algorithm is a two steps growth in which, characteristics are branched into clusters by means of using graph-theoretic clustering means. In the subsequent step, the mainly used representative feature that is robustly affiliated to target classes is peculiar from each cluster to complex the final subset of features. Features in corrected clusters are comparatively sovereign; the clustering-based Scheme of FAST has a high hazard of producing a subset of practical and independent characteristics. In our projected FAST algorithm, it entails the domicile of the minimum spanning tree from a subjective broad graph; the divorce of the minimum spanning tree into a forest by means of every tree bespeak a cluster and the collection of representative appearance from the clusters.

Key words: Redundant features, superfluous characteristics, clusters, FAST, spanning tree

INTRODUCTION

Background Data mining, the eradication of hidden predictive information from large databases is a powerful new technology with great potential to help community focus on the most important instruction in their data warehouses. Data mining tools anticipate future trends and attitude, allowing businesses to make anxious, knowledge-driven decisions. The automated, eventual analyses offered by data mining move above the reasoning of past events administer by reflective tools typical of decision backing systems. Data mining tools can answer business inquiry that traditionally were too time exhausting to resolve. They scour databases for hidden decoration, finding auguring information that authority may miss because it lies outside their chance. Data mining capability are the result of a long process of inquiry and product development. This change commence when business data was first stored on computers, endure with improvements in data entry and more recently, develop technologies that allow users to cross through their data in real time. Data mining takes this evolutionary growth beyond retrospective data access and shipping to prospective and dedicated information distribution. Data mining is ready for application in the business company because it is promoted by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

The scope of data mining: Data mining assume its name from the closeness between searching for valuable business information in a large database for example, finding linked brand in gigabytes of store scanner data and mining a mountain for a vein of profitable ore. Both processes lack either sifting through an endless amount of earthly or intelligently probing it to find absolutely where the value resides. Given databases of ample size and quality, data mining technology can engender new business opportunities by afford these capabilities: Automated prediction of trends and demeanor. Data mining automates the measure of finding predictive

Corresponding Auhor: T. Divya, Department of CSE, KL University, Green fields, Vaddeswaram, Guntur, 522502 Andhra Pradesh, India

intelligence in large databases. Questions that traditionally required expanded hands-on analysis can now be answered precisely from the data quickly. A typical example of a predictive problem is design marketing. Data mining uses data on past promoting mailings to identify the targets most likely to enlarge return on investment in future expressing. Other predictive problems combine forecasting bankruptcy and other forms of default and analyze segments of a population likely to respond equivalently to given events. Automated discovery of previously unknown Patterns. Data mining tools sweep through databases and classify previously hidden patterning one step. An example of diagram discovery is the reasoning of retail sales data to determine seemingly unrelated device that are often purchased together. Other decoration discovery dilemma includes detecting deceitful credit card transactions and identifying divergent data that could perform data entry keying errors.

Feature subset selection is an effective way for dimensionality reduction, expulsion of inappropriate data, rising learning accurateness and mending result unambiguousness. Numerous feature subset selection design have been planned and studied for machine learning applications (Almuallim and Dietterich, 1994). They can be detached into four major league such as: the Wrapper, Embedded and Filter and Hybrid methods. In particular we acquire the minimum spanning tree based clustering algorithms for the reasoning that they do not create that data points are clustered around centers or separated by means of a normal graphic curve and have been greatly used in tradition (Dash and Liu, 1997). The projected feature subdivision selection algorithm FAST was tested and the investigational results demonstrate that, appraise with other disparate types of feature subset selection algorithms, the projected algorithm not only contraction the number of features but also proposal the performances of the acclaimed various types of classifiers (Battiti, 1994). Feature subset excerpt can be analyzed as the process of recognizing and eradicate as many irrelevant and redundant features as promising since: inappropriate features do not put in to the predictive exactitude and redundant characteristics do not redound to getting an appreciate predictor for that they make available mainly advice which is by now present in earlier feature (Butterworth *et al*., 2005; Chikhi and Benhammada, 2009). Based on the minimum spanning tree scheme, we advocate a FAST algorithm which is a two steps action in which; characteristics are divided into clusters by means of using graph-theoretic clustering means (Almuallim and Dietterich, 1994). In the ensuing step, the mainly used representative feature that is robustly pertinent to target classes is distinct from each cluster to complex the final subset of features. Features in corrected clusters are similarly autonomous; the clustering based blueprint of FAST has a high opportunity of producing a subset of effective and independent characteristics

MATERIALS AND METHODS

Inappropriate appearance, all along with redundant confession, severely has an effect on the exactness of the learning machines. Consequently, feature subset collection should be able to admit and take away as much of the unrelated and redundant information as apparent. In addition, preferable feature subsets encase features awfully linked with the class, so far uncorrelated with each other (Bell and Wang, 2000). We build up a novel algorithm shown in Fig. 1 which can capably and effortlessly deal with both inappropriate and redundant characteristics and get hold of a superior feature subset. We obtain this all the way through a novel peculiar selection construction which relaxed of the two associated factor of removal of immaterial features and expulsion of inordinate feature (Dash and Liu, 1997). The earlier achieve features relevant to the target notion by means of removing incorrect ones and the concluding removes inordinate characteristics from applicable ones by means of promote representatives from discrete feature clusters and consequently produces the closing subset (Baker and McCallum, 1998; Chikhi and Benhammada, 2009). The removal of extraneous feature is uncomplicated formerly the right significance assess is delineate, although the rejection of redundant feature is a bit of complicated. In our envisage FAST algorithm, it involve the building of the minimum spanning tree from a subjective broad graph; the separation of the minimal spanning tree into a forest by means of every tree signifying a cluster; and the selection of representative features from the clusters. For the most part of the information contained in inordinate characteristics is by now present in other characteristics. Consequently, redundant features do not add to getting better interpreting capability to the target idea. In order to more exactly initiate the algorithm and for the reason that our projected feature subset collection structure commit inappropriate feature discharge and redundant feature elimination. Appropriate features have tough interaction with target idea so are continually needed for a bst subset whereas superfluous characteristics are not since their values are perfectly concurrent with each other (Das and Liu, 2003). Consequently, notions of feature repetition and feature significance are regularly in terms of feature union and feature target concept association. Mutual

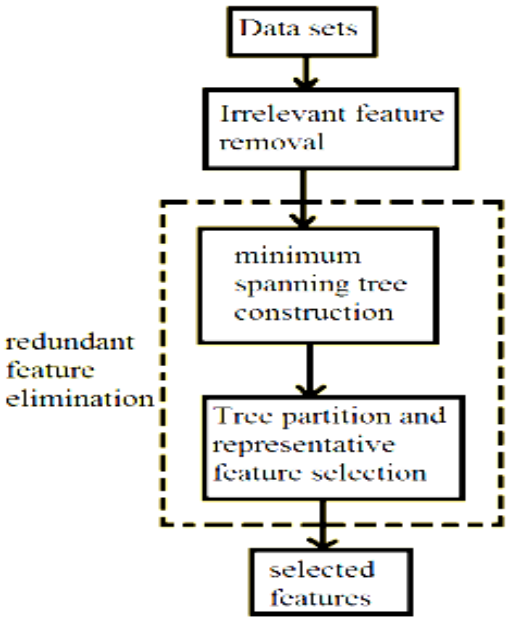

Fig. 1: An overview of feature subset selection algorithum

advice computes how much the allocation of the article values and target classes are at argument from statistical freedom (Biesiada and Duch, 2007). This is a nonlinear inference of company among feature beliefs and target classes

RESULTS AND DISCUSSION

FAST executes excessively well on the microarray data. The confession lies in both the features of the data set itself and the estate of the envisage algorithm. Microarray data has the situation of the large number of characteristics other than small fragment size which can cause curse of dimensionality. In the existence of numerous features, researchers become attentive of that it is broad that a enormous number of characteristics are not illuminating because they are moreover irrelevant or superfluous with deference to the class approach. Consequently, choosing a small number of discriminative heredity from numerous genes is decisive for booming sample categorization. Our projected FAST handily filters out a mass of irrelevant features in the initial step which diminish the likelihood of inappropriately deliver the inappropriate features into the achieve analysis. Then, in the subsequent step, FAST disqualify a large number of outmoded features by means of deciding a single representative characteristic from each cluster of outmoded features. Consequently, only a very small number of discerning characteristics are selected.

CONCLUSION

Based on the minimum spanning tree method, we confirm a FAST algorithm. The algorithm is a two steps action in which, characteristics are divided into bundle by means of using graph-theoretic clustering means. Feature subset selection can be analyzed as the process of recognizing and eliminating as many irrelevant and redundant features as auspicious since: inappropriate features do not put in to the predictive exactitude and redundant characteristics do not react to getting an enhanced forecaster for that they make applicable mainly intelligence which by now present in previous character. In the subsequent step, the mainly used representative character that is robustly pertinent to target classes is distinct from each cluster to complex the final subset of features. Features in corrected clusters are analogously autonomous; the clustering based scheme of FAST has a high prospect of producing a subset of effective and independent characteristics. Estimated FAST algorithm, it entails the domicile of the minimum spanning tree from a subjective general graph; separation of the littlest spanning tree into a forest by equipment of every tree signifying a cluster and assortment of representative character from clusters. Projected character subset selection algorithm FAST was approved and investigational results demonstrate, appraise with other various types of feature subgroup selection algorithms, the projected algorithm not only contraction the number of features but, also approach the performances of the renowned assorted types of classifiers.

REFERENCES

Almuallim, H. and T.G. Dietterich, 1994. Learning boolean concepts in the presence of many irrelevant features. Artif. Intell., 69: 279-305.

Baker, L.D. and A.K. McCallum, 1998. Distributional clustering of words for text classification. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28, 1998, ACM, New York, USA., ISBN:1-58113-015-5, pp: 96-103.

Battiti, R., 1994. Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Networks, 5: 537-550.

Bell, D.A. and H. Wang, 2000. A formalism for relevance and its application in feature subset selection. Mach. Learn, 41: 175-195.

Biesiada, J. and W. Duch, 2007. Feature Selection for High-Dimensional Data: A Pearson Redundancy Based Filter. In: Computer Recognition Systems 2. Kurzynski, M., E. Puchala, M. Wozniak and A. Zolnierek (Eds.). Springer Berlin Heidelberg, Berlin, Germany, ISBN: 978-3-540-75175-5, pp: 242-249.

Butterworth, R., G.P. Shapiro and D.A. Simovici, 2005. On feature selection through clustering. ICDM., 5: 581-584.

Chikhi, S. and S. Benhammada, 2009. ReliefMSS: A variation on a feature ranking ReliefF algorithm. Int. J. Bus. Intell. Data Min., 4: 375-390.

Dash, M. and H. Liu, 1997. Feature selection for classification. Intell. Data Anal., 1: 131-156.

Dash, M. and H. Liu, 2003. Consistency-based search in feature selection. Artif. Intell., 151: 155-176.