A Review an Analysis on Data Reduction and Classification Technique

[1]Ronak Jain, [2]Prof. Sanjay Bansal

Abstract: **In general data mining is the process of data analyzing from different view and summarizing it into useful information. Data mining provides and barring the reference amongst of analytical tools for analyzing data. It allows users to analyze hint alien strange alternative admiration, categorize it, and summarize the relationships identified. The main aim of this paper is to analyses data reduction and classification techniques so that better feature selection features can be found out. Data reduction improves the classification performance in terms of speed, accuracy so the objective of this paper is in the direction of studying reduced data dimensionality methods and analyses their approaches which can be able to provide a high accuracy.**

Keywords: **Data Reduction, Classification, Accuracy, Data Mining.**

1. Introduction

Data mining is team a few of the overpower symbol croak review issues in data mining, which was introduced by Agrawal and Srikant [1] and can be designated as follows: we are disposed a traditional of data-sequences, as the input data. Eternally data combination contains a straightforward list of relationships (item sets). Given a user-specified minimum support threshold (minsup), Progressive pattern mining finds on all sides of subsequences with reference to frequencies more intelligent than minsup in a sequence database.

Various previously studies on connection critique demonstrate focus using only minsup as single prevalence threshold does not excess on the nature of as a matter of actual fact in real-life applications [2]. That is, single minsup implicitly assumes that all items in the database have similar frequency. However, some items may appear very frequently in the database while others rarely appear [3]. Under such circumstance, if we routine the reckoning of minsup counting up snotty, we firmness call for snare those order involving rare items in the database [4].
On the other hand, if we habituated wander give as well pinchbeck, it stability tote a prominent amount of meaningless patterns. Therefore, Liu et al. [3] first address this dilemma (also known as the rare item problem) and propose the concept of multiple minimum supports (MMS in short) to redefine the problem of association rule mining. The method allows users to specify different minsups for different items (MIS values) to reflect their unique natures and generate different association rules, depending on which items are in the patterns.

Some Data Mining algorithms are steadfast to the detection of such captivating subgroups [5]. Anyhow, interesting subgroups are an accepting intermediation of capturing awareness close by the database, as a remedy for by understandability they unequalled label parts of the database [6]. Most suitable algorithms will history repugnance interesting subgroups not as the end deliberation, but as mere edifice blocks for unconcealed descriptions of the genuine regularities [7]. The structures digress are the direction of such algorithms are alike as models, and the existent skirmish of in view of subgroups and zealously fib an absolute avoid of the data is therefore often referred to as modeling[8]. We can take on oneself of the database as a amassing of raw harmony with reference to a painstaking domain. Each insigne serves as a wrapper of the rules that stool this domain [9]. The hew that is induced non-native the raw data is a shortened affirmation of the action of the domain, unique the statistics of kin. Having a model allows us to convince about the domain, for example to see causes for diseases in transferable databases of patients [10-12]. Roughly evidently data Mining is often utilitarian in order to derive prophetic models [13]. If we receive that the database lower than enumeration is but a facsimile of a larger or evolvement populace of individuals, we can justify the induced model to reckon on the behavior of new individuals. We seat secure a sculpt reading in any case the answer depends on alternate categorize of the clientele, in all directions the wish of predicting despite that other patrons will respond to the offer[13]. A total of discretion and relevancy underpinning estimation be saved by desolate around customers with a predicted interest [14].

2. LITERATURE REVIEW

In 2010, Marcano-Cedeño et al. [15] suggested that the feature selection has been widely used to reduce the data dimensionality. Information lessening enhances the grouping execution, the estimation capacity, and example acknowledgment frameworks as far as rate, exactness and straightforwardness. A technique to lessen the quantity of components in neighborhood pursuit is the consecutive look calculations. They have exhibited a component choice method based on Sequential Forward Selection (SFS) and Feed Forward Neural Network (FFNN) to estimate the prediction error as a selection criterion. Three well-known database have been used to test the SFS-FFNN with Artificial Metaplasticity on Perceptron Multilayer (AMMLP). The AMMLP is another technique connected for arrangement of examples. The outcomes acquired by SFS-FFNN with AMMLP in arrangement exactness are predominant than got by routine BP calculation and other later highlight choice calculations connected to the same database. By these reasons the proposed technique SFS-FFNN with AMMLP is a fascinating contrasting option to decrease the information dimensionality and give a high exactness.

In 2013, Naseriparsa et al. [16] proposed a hybrid feature selection method which takes advantages of wrapper subset evaluation with a lower cost and improves the performance of a group of classifiers. The technique utilizes mix of test area sifting and resampling to refine the example space and two component subset assessment strategies to choose solid elements. This strategy uses both element space and test area in two stages. The principal stage channels and resamples the specimen area and the second stage receives a half breed strategy by data pick up, wrapper subset assessment and hereditary hunt to locate the ideal component space. Tests completed on various sorts of datasets from UCI Repository of Machine Learning databases and the outcomes demonstrate an ascent in the normal execution of five classifiers (Naïve Bayes, Logistic, Multilayer Perceptron, Best First Decision Tree and JRIP) all the while and the arrangement mistake for these classifiers diminishes significantly. The trials likewise demonstrate that this strategy outflanks other element determination strategies with a lower cost.

In 2013, Ozarkar et al. [17] suggested that the spam is a key problem in electronic communication, including large-scale email systems and the growing number of blogs. According to the author there are several research work are going in the automatic detection of spam emails using classification techniques such as SVM, NB, MLP, KNN, ID3, J48,Random Tree, etc. For spam dataset it is conceivable to have substantial number of preparing occasions. Based on this, they have made utilization of Random Forest and Partial Decision Trees calculations to arrange spam versus non-spam messages. These calculations outflanked the already executed calculations as far as precision and time many-sided quality. As a preprocessing step we have utilized element choice strategies, for example, Chi-square, Information pick up, Gain proportion, Symmetrical instability and Correlation. This permitted us to choose subset of important, non-excess and generally contributing components to have an additional advantage as far as impromptu creation in precision and decreased time unpredictability.

In 2014, Murthy et al. [18] suggested Microarray Gene Profile for assessing the global patterns of thousands of genes under different varying conditions. It provides important insights about the underlying genetic causes for diseases, ultimately allowing the development of modern chemical entities as medical-kit drug candidates. The informatics analysis and integration of microarray gene expression pattern are difficult for understanding or interpretation of gene array features. In this paper, we discuss the deterministic computational analysis of: the identification of differentially expressed genes using statistical methods, the discovery of gene clusters, and the classification of biological samples using standard clustering and classification approaches.

In 2014, Kumar et al. [19] suggested that the data mining provides an automatic extraction of useful and relevant content, often previously unknown information from large databases or data sets. Many data mining applications contain high dimensional data. The High dimensionality diminishes the execution of the mining calculations and builds the time and space required for preparing the information. The high dimensionality issue is determined utilizing the Dimensionality Reduction (DR) system. The DR is partitioned into two: component determination and highlight extraction. They have used a subtle element study has been done to know how the dimensionality issue has settled by utilizing the two diverse systems. Furthermore different factual measures are disclosed to choose the most significant components and diverse measurable procedures are broke down to remove the new arrangement of highlights frame the first components.

In 2015, Vaska et al. [20] suggested that the E-Health has grown popular due to a wide range of services provided. The part of a patient has likewise changed in today's social insurance as they are relied upon to utilize ICT administrations to pick up data and information to think about their prosperity. In the field of information mining grouping is a broadly utilized system for finding designs as a part of hidden information. Customary grouping calculations are ordinarily constrained to taking care of datasets that contain either numeric or unmitigated properties. Nonetheless, datasets with blended sorts of qualities are likewise normal, in actuality, information mining applications. They have presented a cluster feature based incremental clustering algorithm; MCIFA (Cluster Feature-Based Incremental Clustering Approach to mixed data) is applied on the diabetes dataset to check its suitability in the medical domain. The accomplished grouping precision in results area demonstrates this is in fact reasonable for medicinal space and can be utilized for 'e-endorsing'. However, it should be adjusted in order to expand the bunching precision as the rate of permitted mistake rate in therapeutic area ought to be as little as could reasonably be expected.

In 2015, Chakraborty et al. [21] presented a feature selection method based on a multilayer perceptron (MLP) neural network, called feature selection MLP (FSMLP). They have explained how FSMLP can select essential features and discard derogatory and indifferent features. They showed the viability of the calculations utilizing a few information sets including a manufactured information set. They additionally demonstrate that the chose elements are satisfactory to take care of the current issue. Here, they have considered a measure of direct reliance to control the repetition. The utilization of nonlinear measures of reliance, for example, shared data, is direct. Here, there are some focal points of the proposed plans. They don't require express assessment of the component subsets. Here, component choice is incorporated into outlining of the basic leadership framework. Thus, it can take a gander at all elements together and get whatever is important. Their techniques can represent conceivable nonlinear unpretentious connections between components, and also that between elements, apparatuses, and the issue being explained. They can likewise control the level of repetition in the chose highlights. Of the two learning plans, mFSMLP-CoR, enhances the execution of the framework, as well as altogether decreases the reliance of the system's conduct on the introduction of association weights.

In 2015, Senthilkumar et al. [22] suggested that the health care data are having exponential growth in volume and complexity. They have presented a correlation attribute evaluation feature selection for dimensionality reduction and t-test for comparing the performance of different classifiers before and after dimensionality reduction. Distinctive classifiers utilized as a part of this work are Naïve Bayes (NB), k- Nearest Neighbour (kNN), Classification tree (CT) and Clark and Nilbert2 (CN2). Empirical results shown that CN2 classifier is best for the multi-dimensional thyroid dataset by comparing classification accuracy of the different classifiers. There is no significant difference between before and after the dimensionality reduction in the four classifiers in the performance measure.

In 2016, Baz et al. [23] suggested that the common aim of all our daily activities is providing services to others or ourselves. They have addressed the gap in improving the quality of their customer service, and enhance the existing services by proposing a number of data analysis techniques that can be utilized to improve the quality of customer service and enhance the existing services in the Saudi government sector. They have introduced a relational database structure that can be utilized to apply the proposed data analysis techniques.

In 2016, Xu et al. [24] suggested that the dimension reduction is an important in pattern analysis and machine learning, and it has wide applications in feature representation and pattern classification. They also suggested that the, sliced inverse regression (SIR) has pulled in much research endeavors because of its viability and adequacy in measurement diminishment. Be that as it may, two downsides restrict further uses of SIR. To start with, the calculation many-sided quality of SIR is typically high in the circumstance of high-dimensional information. Second, sparsity of projection subspace is not all around dug for enhancing the component determination and model translation capacities. They have proposed to compute the SIR projection vectors in the spectral space, at that point an approximated relapse arrangement can be gotten with a quicker speed. In addition, the versatile tether is utilized to achieve a scanty and universally ideal arrangement, which is essential in variable determination. To finish the vigorous example arrangement assignment with defilements, an entropy-based and class-wise relapse model is planned in this paper. It takes a smooth punishment rather than sparsity requirement in the relapse coefficients, and it can be led in class-wise, along these lines it is more adaptable by and by. Broad tests are led by utilizing some genuine and benchmark information sets, e.g., high-dimensional facial pictures and quality microarray information, to assess the new calculations. The new recommendations accomplish focused results and are contrasted and other best in class strategies.

3. PROBLEM DOMAIN

There are several works are already progress in this direction to. Some of the gaps identified are following:

1) The observations based on the data are known to desire to a prespecified class and the main aim is to design or develop the predictors for the new observations to these classes or labels but the staring labels is the predetermination of the whole so the selection criteria is always important.
2) The data behavior and nature are different so any static method will not help in determine the performance.
3) A strategy to reduce the data dimension without affecting the whole classification is needed.
4) The attribute reduction can be qualified with fewer errors and does not affect the input and target variants.

4. ANALYSIS

Our analysis based on the above study suggests the following in table 1:

Table 1 Comparison of various Dimension reduction techniques

S.No	Method	Working Principle
1	SVD	They are applied for the identification and extraction of structural data. It is factorization method.
2	PCA	They are applied for feature extraction and able to convert linear combinations.
3	LDA	They are mainly used for small size sample for feature transformation and selection. It is a classification based approach.
4	SVM	It is a supervised algorithm applied on large number of feature vectors like gene expression data.
5	Independent Component Analysis(ICA)	It is an unsupervised algorithm applied on covariance matrix and use in separation of multivariate source.
6	Canonical Correlation Analysis (CCA)	It is an unsupervised algorithm used for DNA micro array expressions used in discovering linear combinations.
7	Neural Network (NN)	It is used to solve classification problems which can be used to train, validate and test.

5. CONCLUSIONS

This paper addresses dimensionality reduction issues in order for both high-dimensional multivariate and utilitarian information. High-dimensional information alludes to information with countless, frequently bigger than the number of perceptions. High-dimensional information is experienced in an extensive variety of ranges, for example, building, biometrics, psychometrics, and neuroimaging. Characterizing this information is a troublesome issue in

light of the fact that the tremendous number of variables stances difficulties to customary arrangement strategies and renders numerous traditional methods unreasonable. A characteristic arrangement is to include a dimensionality decrease venture before a grouping method is connected. As this dissertation main aim is to prove that all the attributes are not always affect the results so the less attributes results are also same validity and proved to be more efficient. It will also improve the classification performance in terms of speed, accuracy.

REFERENCES

[1] R. Agrawal and R. Srikant, "Mining sequential patterns," Data Engineering (ICDE'95), Taipei, Taiwan, Mar, 1995, pp.3-14.

[2] Y.-H. Hu and Y.-L. Chen, "Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism," Decision Support Systems, vol. 42, is. 1, pp.1-24, 2006.

[3] Shrivastava P, Gupta H. A Review of Density-Based clustering in Spatial Data. International Journal of Advanced Computer Research (IJACR). 2012 Sep;2.

[4] B. Liu, W. Hsu, and Y. Ma, "Mining association rules with multiple minimum supports,", Proceedings of the fifth ACM SIGKDD international conference, San Diego, CA, USA August 15-18, 1999,p.341.

[5] Y. Lee, T. Hong, and W. Lin, "Mining fuzzy association rules with multiple minimum supports using maximum constraints," Lecture notes in computer science, vol. 3014, pp.1283-1290, 2004.

[6] Amit Kumar Sinha and Vinay Singh, " Transformation of LOG file using LIPT technique " , International Journal of Advanced Computer Research (IJACR), Volume-6, Issue-23, March-2016 ,pp.58-64.

[7] C. Ezeife and Y. Lu, "Mining web log sequential patterns with position coded pre-order linked wap-tree," Data Mining and Knowledge Discovery, vol. 10, no. 1, pp.5-38, 2005.

[8] Khare P, Gupta H. Finding frequent pattern with transaction and occurrences based on density minimum support distribution. International Journal of Advanced Computer Research (IJACR). 2012 Sep;2(3):5.

[9] T. Hey and A. Trefethen, "Cyberinfrastructure for e-science," Science Magazine, vol. 308, no. 5723, pp. 817–821, 2005.

[10] J. Gray, D. T. Liu, M. Nieto-Santisteban, A. Szalay, D. J. DeWitt, and G. Heber, "Scientific data management in the coming decade," SIGMOD Rec., vol. 34, no. 4, pp. 34–41, 2005.

[11] Dubey AK, Dubey AK, Agarwal V, Khandagre Y. Knowledge discovery with a subset-superset approach for Mining Heterogeneous Data with dynamic support. InSoftware Engineering (CONSEG), 2012 CSI Sixth International Conference on 2012 Sep 5 (pp. 1-6). IEEE.

[12] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams: A review," ACM SIGMOD Record, vol. Vol. 34,no. 1, 2005.

[13] C. C. Aggarwal, Data Streams: models and algorithms. Springer, 2007.

[14] J. Han, and M. Kamber, Data Mining: Concepts and Techniques,Morgan Kaufmann, 2001.

[15] Marcano-Cedeño A, Quintanilla-Domínguez J, Cortina-Januchs MG, Andina D. Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network. InIECON 2010-36th Annual Conference on IEEE Industrial Electronics Society 2010 Nov 7 (pp. 2845-2850). IEEE.

[16] Naseriparsa M, Bidgoli AM, Varaee T. A Hybrid Feature Selection method to improve performance of a group of classification algorithms. arXiv preprint arXiv:1403.2372. 2014 Mar 8.

[17] Ozarkar P, Patwardhan M. Efficient spam classification by appropriate feature selection. Global J. Comput. Sci. Technol. Softw. Data Eng. 2013; 13(5).

[18] Murthy VB, Varma GP. Deterministic Models for Microarray Gene Expression Profiles. International Journal of Advanced Computer Research. 2014 Jun 1; 4(2):459.

[19] V. Arul Kumar, N. Elavarasan, "A Survey on Dimensionality Reduction Technique", International Journal of Emerging Trends & Technology in Computer Science, Volume 3, Issue 6, November-December 2014.

[20] Vaska JS, Sowjanya AM. Clustering Diabetics Data Using M-CFICA. International Journal of Advanced Computer Research. 2015 Sep 1;5(20):327.

[21] Chakraborty R, Pal NR. Feature selection using a neural framework with controlled redundancy. Neural Networks and Learning Systems, IEEE Transactions on. 2015 Jan;26(1):35-50.

[22] Senthilkumar D, Sheelarani N, Paulraj S. Classification of multi-dimensional thyroid dataset using data mining techniques: comparison study. Advances in Natural and Applied Sciences. 2015 Jun 1;9(6 SE):24-9.

[23] Baz A. Efficient data analysis approaches to enhance the quality of customer service in Saudi Government sector. International Journal of Advanced Computer Research (IJACR), Volume-6, Issue-22, January-2016, pp. 25-30.

[24] X. L. Xu; C. X. Ren; R. C. Wu; H. Yan, "Sliced Inverse Regression With Adaptive Spectral Sparsity for Dimension Reduction," in IEEE Transactions on Cybernetics , vol.PP, no.99, pp.1-13.