World Journal of Engineering Research and Technology

WJERT

www.wjert.org

SJIF Impact Factor: 5.924

A REVIEW ON DIMENSIONALITY REDUCTION TECHNIQUES IN DATA MINING

Wasim Akram* and Sriram Yadav

[1]M. Tech. Scholar, CSE, MITS Bhopal.

[2]A.P., CSE, MITS Bhopal.

Article Received on 20/09/2019 **Article Revised on 10/10/2019** **Article Accepted on 31/10/2019**

***Corresponding Author**

Wasim Akram

M. Tech. Scholar, CSE,

MITS Bhopal.

ABSTRACT

Data mining is a form of knowledge discovery essential for solving problems in a specific domain. Classification is a technique used for discovering classes of unknown data. Various methods for classification exists like Bayesian, Decision Trees and Rule based neural networks etc. Before applying any mining technique, irrelevantattributes needs to be filtered. Filtering is done using different feature selection techniques like wrapper, filter, and embedded technique. Feature selection plays an important role in data mining and machine learning. It helps to reduce the dimensionality of data and increase the performance of classification algorithms. A variety of feature selection methods have been presented in state-of-the-art literature to resolve feature selection problems such as large search space in high dimensional datasets like in microarray. However, it is a challenging task to identify the best feature selection method that suits a specific scenario or situation. Dimensionality reduction in data mining focuses on representing data with minimum number of dimensions such that its properties are not lost and hence reducing the underlying complexity in processing the data. Principal Component Analysis (PCA) is one of the prominent dimensionality reduction techniques widely used in network traffic analysis.

KEYWORDS: Feature selection, Dimensionality reduction, Classification, Data mining, Machine learning, Neural Networks, Decision trees and PCA.

I. INTRODUCTION

Data mining is a step in the whole process of knowledge discovery which can be explained as a process of extracting or mining knowledge from large amounts of data.[1] Data mining is a form of knowledge discovery essential for solving problems in a specific domain. Data mining can also be explained as the non trivial process that automatically collects the useful hidden information from the data and is taken on as forms of rule, concept, pattern and so on.[2] The knowledge extracted from data mining, allows the user to find interesting patterns and regularities deeply buried in the data to help in the process of decision making. The data mining tasks can be broadly classified in two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions. According to different goals, the mining task can be mainly divided into four types: class/concept description, association analysis, classification or prediction and clustering analysis.[3]

Dimensionality reduction is the most important and popular technique to eliminate irrelevant and redundant features from the datasets. It can be categorized mainly into two sub-categories i.e. feature extraction and feature selection.[5] The feature extraction approach merges multiple features to compose a new feature with lower feature space. Examples of feature extraction methods are Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA) and Linear Discriminant Analysis (LDA).[6] On the other hand, feature selection approach selects a subset of features from the dataset and aims to minimize feature redundancy and maximize the feature relevance to the target class label. Some examples of feature selection techniques are Chi-square[7], Fisher score[8], Information Gain[9], ReliefF[5] and minimum redundancy and maximum relevance (mRmR).[10]

Both the techniques i.e. feature extraction and feature selection can improve the learning performance in terms of accuracy, model interpretability, computational complexity and storage requirements. Feature selection is considered superior than feature extraction considering interpretability and readability. Maintaining the original features in the subset resulting from feature selection has a great significance in different areas of research, for instance, identifying the relevant genes to target disease in medical domain.[11]

II. Data Preprocessing

Data available for mining is raw data. Data may be in different formats as it comes from different sources, it may consist of noisy data, irrelevant attributes, missing data etc. Data needs to be preprocessed before applying any kind of data mining algorithm which is done using following steps.[12]

Data Integration – If the data to be mined comes from several different sources data needs to be integrated which involves removing inconsistencies in names of attributes or attribute value names between data sets of different sources.

Data Cleaning –This step may involve detecting and correcting errors in the data, filling in missing values, etc. Some data cleaning methods are discussed in.[13,14]

Discretization –When the data mining algorithm cannot cope with continuous attributes, discretization needs to be applied. This step consists of transforming a continuous attribute into a categorical attribute, taking only a few discrete values. Discretization often improves the comprehensibility of the discovered knowledge.[15,16]

Attribute Selection – Not all attributes are relevant so for selecting a subset of attributes relevant for mining, among all original attributes, attribute selection is required.

III. Feature Selection

The selection of optimal features adds an extra layer of complexity in the modeling as instead of just finding optimal parameters for full set of features, first optimal feature subset is to be found and the model parameters are to be optimized.[17] Attribute selection methods can be broadly divided into filter and wrapper approaches. In the filter approach the attribute selection method is independent of the data mining algorithm to be applied to the selected attributes and assess the relevance of features by looking only at the intrinsic properties of the data. In most cases a feature relevance score is calculated, and low scoring features are removed. The subset of features left after feature removal is presented as input to the classification algorithm. Advantages of filter techniques are that they easily scale to high dimensional datasets are computationally simple and fast, and as the filter approach is independent of the mining algorithm so feature selection needs to be performed only once, and then different classifiers can be evaluated. Disadvantages of filter methods are that they ignore the interaction with the classifier and that most proposed techniques are univariate which means that each feature is considered separately, thereby ignoring feature dependencies, which may lead to worse classification performance when compared to other types of feature selection techniques. In order to overcome the problem of ignoring feature dependencies, a number of multivariate filter techniques were introduced, aiming at the incorporation of feature dependencies to some degree.

Wrapper methods embed the model hypothesis search within the feature subset search. In the wrapper approach the attribute selection method uses the result of the data mining algorithm to determine how good a given attribute subset is. In this setup, a search procedure in the space of possible feature subsets is defined, and various subsets of features are generated and evaluated. The major characteristic of the wrapper approach is that the quality of an attribute subset is directly measured by the performance of the data mining algorithm applied to that attribute subset. The wrapper approach tends to be much slower than the filter approach, as the data mining algorithm is applied to each attribute subset considered by the search. In addition, if several different data mining algorithms are to be applied to the data, the wrapper approach becomes even more computationally expensive.[18] Advantages of wrapper approaches include the interaction between feature subset search and model selection, and the ability to take into account feature dependencies. A common drawback of these techniques is that they have a higher risk of over fitting than filter techniques and are very computationally intensive. Another category of feature selection technique was also introduced, termed embedded technique in which search for an optimal subset of features is built into the classifier construction, and can be seen as a search in the combined space of feature subsets and hypotheses.

Just like wrapper approaches, embedded approaches are thus specific to a given learning algorithm. Embedded methods have the advantage that they include the interaction with the classification model, while at the same time being far less computationally intensive than wrapper methods.[19]

IV. Classification

Feature section techniques can be categorized into supervised.[20,21], semi-supervised[22,23] and unsupervised[24,25] approaches. Supervised feature selection can further be divided into filter, wrapper and embedded models.[26] Filter model aims to select the features independently without considering any learning algorithm.[27] Semi-supervised learning is usually used when a small subset of labeled examples is available, together with a large number of unlabeled examples. Unsupervised method only depends on clustering quality measure[28] and is less constrained search problem having no consideration of class labels. Generally, a feature selection technique consists of four steps[29] i.e. subset feature generation, subset feature evaluation, stopping criterion and result validation.

Dimensionality Reduction

1. Feature Extraction

(i) PCA

(ii) CCA

(iii) LCA

2. Feature Selection

(i) Filter Model

(ii) Wrapper Model

(iii) Embedded Model

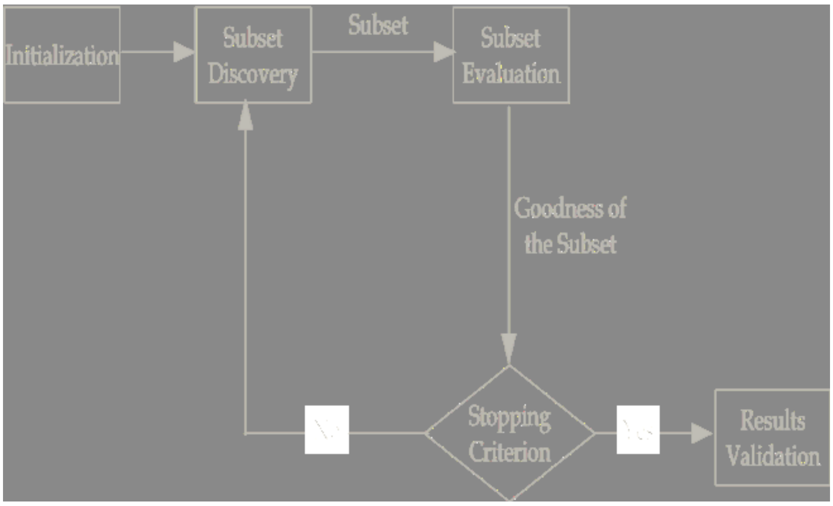

Figure 1: General Framework for Feature Selection.[31]

Classification Techniques

A. Rule Based Classifiers

Rule based classifiers deals with the the discovery of high-level, easy-to-interpret classification rules of the form if-then. The rules are composed of two parts mainly rule antecedent and rule consequent. The rule antecedent, is the if part, specifies a set of conditions referring to predictor attribute values, and the rule consequent, the then part, specifies the class predicted by the rule for any example that satisfies the conditions in the rule antecedent. These rules can be generated using different classification algorithms, the most well known being the decision tree induction algorithms and sequential covering rule induction algorithms.[30]

B. Bayesian Networks

A Bayesian network (BN) consists of a directed, acyclic graph and a probability distribution for each node in that graph given its immediate predecessors.[31] A Bayes Network Classifier is based on a Bayesian network which represents a joint probability distribution over a set of categorical attributes. It consists of two parts, the directed acyclic graph G consisting of nodes and arcs and the conditional probability tables. The nodes represent attributes whereas the arcs indicate direct dependencies. The density of the arcs in a BN is ne measure of its complexity. Sparse BNs can represent simple probabilistic models (e.g., naïve Bayes models and hidden Markov models), whereas dense BNs can capture highly complex models. Thus, BNs provide a flexible method for probabilistic modeling.[32]

C. Decision Tree

A Decision Tree Classifier consists of a decision tree generated on the basis of instances. The decision tree has two types of nodes: a) the root and the internal nodes, b) the leaf nodes. The root and the internal nodes are associated with attributes, leaf nodes are associated with classes. Basically, each non-leaf node has an outgoing branch for each possible value of the attribute associated with the node. To determine the class for a new instance using a decision tree, beginning with the root, successive internal nodes are visited until a leaf node is reached. At the root node and at each internal node, a test is applied. The outcome of the test determines the branch traversed, and the next node visited. The class for the instance is the class of the final leaf node.[33]

D. Nearest Neighbour

A Nearest Neighbor Classifier assumes all instances correspond to points in the n-dimensional space. During learning, all instances are remembered. When a new point is classified, the k-nearest points to the new point are found and are used with a weight for determining the class value of the new point. For the sake of increasing accuracy, greater weights are given to closer points.[17]

E. Artificial Neural Network

An artificial neural network, often just called a neural network is a mathematical model or computational model based on biological neural networks, in other words, is an emulation of biological neural system. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.[34] A Neural Network Classifier is based on neural networks consisting of interconnected neurons. From a simplified perspective, a neuron takes positive and negative stimuli (numerical values) from other neurons and when the weighted sum of the stimuli is greater than a given threshold value, it activates itself. The output value of the neuron is usually a non-linear transformation of the sum of stimuli. In more advanced models, the non-linear transformation is adapted by some continuous functions.

F. Support vector machines

Support Vector Machines[35] are basically binary classification algorithms. Support Vector Machines (SVM) is a classification system derived from statistical learning theory. It has been applied successfully in fields such as text categorization, hand-written character recognition, image classification, bio-sequences analysis, etc. The SVM separates the classes with a decision surface that maximizes the margin between the classes. The surface is often called the optimal hyperplane, and the data points closest to the hyperplane are called support vectors. The support vectors are the critical elements of the training set. The mechanism that defines the mapping process is called the kernel function. The SVM can be adapted to become a nonlinear classifier through the use of nonlinear kernels. SVM can function as a multiclass classifier by combining several binary SVM classifiers. The output of SVM classification is the decision values of each pixel for each class, which are used for probability estimates. The probability values represent "true" probability in the sense that each probability falls in the range of 0 to 1, and the sum of these values for each pixel equals 1. Classification is then performed by selecting the highest probability. SVM includes a penalty parameter that allows a certain degree of misclassification, which is particularly important for nonseparable training sets. The penalty parameter controls the trade-off between allowing training errors and forcing rigid margins. It creates a soft margin that permits some misclassifications, such as it allows some training points on the wrong side of the hyperplane. Increasing the value of the penalty parameter increases the cost of misclassifying points and forces the creation of a more accurate model that may not generalize well.[36]

G. Rough Sets

Any set of all indiscernible (similar) objects is called an elementary set. Any union of some elementary sets is referred to as a crisp or precise set - otherwise the set is rough (imprecise, vague). Each rough set has boundary-line cases, i.e., objects which cannot be with certainty classified, by employing the available knowledge, as members of the set or its complement. Obviously rough sets, in contrast to precise sets, cannot be characterized in terms of information about their elements. With any rough set a pair of precise sets - called the lower and the upper approximation of the rough set is associated. The lower approximation consists of all objects which surely belong to the set and the upper approximation contains all objects which possible belong to the set. The difference between the upper and the lower approximation constitutes the boundary region of the rough set. Rough set approach to data analysis has many important advantages like provides efficient algorithms for finding hidden patterns in data, identifies relationships that would not be found using statistical methods, allows both qualitative and quantitative data, finds minimal sets of data (data reduction), evaluates significance of data, easy to understand.[37]

H. Fuzzy Logic

Fuzzy logic is a multivalued logic different from "crisp logic", where binary sets have two valued logic. Fuzzy logic variables have truth value in the range between 0 and 1. Fuzzy logic is a superset of conventional Boolean logic that has been extended to handle the concept of partial truth. A membership function (MF) is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. Fuzzy Logic consists of Type 1 and Type 2 fuzzy logic. Type 1 fuzzy contains the constant values. A Type-2 Fuzzy Logic is an extension of Type 1 Fuzzy Logic in which the fuzzy sets comes from Existing Type 1 Fuzzy. A type-2 fuzzy set contains the grades of membership that are themselves fuzzy. A Type-2 membership grade can be any subset in the primary membership. For each primary membership there exists a secondary membership that defines the possibilities for the primary membership. Type-1 Fuzzy Logic is unable to handle rule uncertainties. Type-2 Fuzzy Logic can handle rule uncertainties effectively and efficiently.[38] Type 2 Fuzzy sets are again characterized by IF–THEN rules.[39] Type-2 Fuzzy is computationally intensive because type reduction is very intensive. Type-2 fuzzy is used for modeling uncertainty and imprecision in a better way. The type-2 fuzzy sets are called as “fuzzy fuzzy” sets where the fuzzy degree of membership is fuzzy itself that results from Type 1 Fuzzy.[40]

V. CONCLUSIONS AND FUTURE WORK

High dimensional data produces serious challenges for existing learning algorithms in the fields of data mining and machine learning. Such data may include redundant and irrelevant features which may mislead the learning algorithm and degrade the performance. In order to address the curse of dimensionality, dimensionality reduction techniques i.e. feature extraction and feature selection was introduced in the literature. In this paper, we reviewed different feature selection models i.e. Filter, Wrapper and Embedded models, state-of-the-art algorithms, their types and their complexities are also critically analyzed in each model. The applicability and suitability of different algorithms like univariate and multivariate techniques in different situations is also discussed. The tradeoff among efficiency, accuracy and computational cost of different algorithms with respect to distinct types of data is also presented. The contribution of feature selection methods in different dynamic areas such as microarray analysis, image classification and text categorization is also highlighted. Last but not least the most significant issues and challenges of feature selection methods were also described that identify the future research directions in this area.

REFERENCES

1. Anomaly detection in IP networkswith Principal Component Analysis. In: Communications and Information Technology, IEEE.http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html(18.02.16).Ringberg, H., et al., 2007.
2. B. Pfahringer, “Supervised and unsupervised discretization of continuous features”, Proc. 12th Int. Conf. Machine Learning, 1995; 456-463.
3. Brauckhoff, D., et al., 2009. Applying PCA for traffic anomaly detec-tion: problems and solutions. In: INFOCOM 2009, IEEE. Issariyapat, C., Kensuke, F., 2009.
4. C.W. Hsu, C.C. Chang and C.J. Lin, “A practical guide to support vector classification”,http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, 2003
5. Chapter 2 of book Data Classification, Algorithms and Applications by Charu C. Aggarwal Chapman and Hall/CRC, 2014; 37–64. Print ISBN: 978-1-4665-8674-1 eBook ISBN: 978-1-4665-8675-8
6. D. Pyle, Data preparation for data mining, 1st Vol., Morgan Kaufmann publisher, San Francisco, 1999.
7. Darwiche, Modeling and Reasoning with Bayesian Networks, Cambridge University Press, 2009.

8. E. Simoudis, B. Livezey B and R. Kerber R , “Integrating inductive and deductive reasoning for data mining”, In: Fayyad UM, Piatetsky-Shapiro G, Smyth P and Uthurusamy R. (Eds.) Advances in knowledge discovery and data mining, AAAI/MIT Press, California, 1996; 353-373.
9. G.F. Cooper, P. Hennings-Yeomans, S. Visweswaran and M. Barmada, “An Efficient Bayesian Method for Predicting Clinical Outcomes from Genome-Wide Data”, AMIA 2010 Symposium Proceedings, 2010; 127-131.
10. G.L. Pappa and A.A. Freitas, Automating the Design of Data Mining Algorithms. An Evolutionary Computation Approach, Natural Computing Series, Springer, 2010.
11. Guyon, N. Matic and V. Vapnik, “Discovering informative patterns and data cleaning”, In: Fayyad UM, Piatetsky-Shapiro G, Smyth P and Uthurusamy R. (ed) Advances in knowledge discovery and data mining, AAAI/MIT Press, California, 1996; 181- 203.
12. H. Liu, R. Setiono, Chi2: feature selection and discretization of numeric attributes, in: Tools with Artificial Intelligence, Proceedings., Seventh International Conference on, IEEE, 1995; 388.
13. H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell, 2005; 27(8): 1226–1238.
14. I.H. Witten, E. Frank and M.A. Hall, Data mining practical machine learning tools and techniques, Morgan Kaufmann ublisher, Burlington 2011 [2] J. Han and M. Kamber, Data mining concepts and techniques, Morgan Kaufmann, San Francisco, 2006.
15. J. Catlett, “On changing continuous attributes into ordered discrete attributes”, In Y. Kodratoff (ed), Machine Learning—EWSL-91, Springer-Verlag, New York, 1991; 164-178.
16. J. Weston, A. Elisseff, B. Schoelkopf, and M. Tipping. Use of the zero norm with linear models and kernel methods. Journal of Machine Learning Research, 2003; 3: 1439–1461.
17. J.G. Dy and C.E. Brodley. Feature subset selection and order identification for unsupervised learning. In In Proc. 17th International Conference on Machine Learning, pages 247–254. Morgan Kaufmann, 2000.
18. J.R. Quinlan, Induction of decision trees, Mach. Learn, 1986; 1(1): 81–106.
19. Kononenko, Estimating attributes: analysis and extensions of relief, in: Machine Learning: ECML-94, Springer, 1994; 171–182.
20. L. Song, A. Smola, A. Gretton, K. Borgwardt, and J. Bedo. Supervised feature selection via dependence estimation. In International Conference on Machine Learning, 2007.

21. L. Tari, C. Baral and S. Kim, “Fuzzy c-means clustering with prior biological knowledge”, Journal of Biomedical Informatics, 2009; 42(1): 74-81.
22. M. Garofalakis, D. Hyun, R. Rastogi and K. Shim, “Building Decision Trees with Constraints”, Data Mining and Knowledge Discovery, 2003; 7(2): 187–214.
23. M.L. Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn and A.K. Jain, “Dimensionality Reduction Using Genetic Algorithms”, IEEE Transactions On Evolutionary Computation, 2000; 4: 2.
24. Network intrusiondetection system dataset and its comparison with KDD CUP99Dataset. In: II AH-ICI, IEEE. Zhang, B., et al., 2012. PCA-subspace method — is it good enough fornetwork-wide anomaly detection. In: Network Operations and Management Symposium (NOMS), IEEE.
25. P. Mitra, C. A. Murthy, and S. Pal. Unsupervised feature selection using feature similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002; 24; 301– 312.
26. R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, 1999.
27. Sensitivity of PCA for traffic anomalydetection. ACM SIGMETRICS Perform. Eval. Rev., 35(1). http://www.unb.ca/research/iscx/dataset/iscx-IDS-dataset. html(18.02.16). Vasudevan, A.R., et al., 2011. SSENet-2011:
28. T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. Springer, 2001.
29. T.J. Shan, H. Wei and Q. Yan, “Application of genetic algorithm in data mining”, 1st Int Work Educ Technol Comput Sci, IEEE, 2009; 2: 353- 356Z.Z. Shi, Knowledge discovery, Tsinghua University Press, Beijing, 2001
30. T.M. Mitchell, Machine Learning, McGraw-Hill Companies, USA, 1997.
31. V. Bolon-Canedo, N. Sanchez-Marono, A. Alonso-Betanzos, J.M. Benitez, F. Herrera, “A review of microarray datasets and applied feature selection Methods”, Information Sciences, 2014; 282: 111–135.
32. V. N. Vapnik, Statistical Learning Theory, Wiley New York., 1998.
33. W. Daelemans, V. Hoste, F.D. Meulder and B. Naudts, “Combined Optimization of Feature Selection and Algorithm Parameter Interaction in Machine Learning of Language”, Proceedings of the 14th European Conference on Machine Learning (ECML-2003), Lecture Notes in Computer Science 2837, Springer-Verlag, Cavtat-Dubrovnik, Croatia, 2003; 84-95.

34. Y. Saeys, I. Inza and P. Larranaga, “A review of feature selection techniques in bioinformatics”, Bioinformatics, 2007; 19: 2507–17.
35. Y. Singh Y, A.S. Chauhan, “Neural Networks in Data Mining”, Journal of Theoretical and Applied Information Technology, 2005; 37-42.
36. Z. Pawlak, “Rough sets”, International Journal of Computer and Information Sciences, 1982; 341- 356.
37. Z. Xu, R. Jin, J. Ye, M. Lyu, and I. King. Discriminative semi-supervised feature selection via manifold regularization. In IJCAI" 09: Proceedings of the 21th International Joint Conference on Artificial Intelligence, 2009.
38. Z. Zhao and H. Liu. Semi-supervised feature selection via spectral analysis. In Proceedings of SIAM International Conference on Data Mining, 2007.