Feature Selection for High Dimensional and Imbalanced Data- A Comparative Study

Kokane Vina A., Lomte Archana C.

Abstract: **The recent increase of data poses a severe challenge in data extracting. High dimensional data can contain high degree of irrelevant and redundant information. Feature selection is the process of eliminating irrelevant data set with respect to the task to be performed. Several features selection techniques are used to improve the efficiency of various machine learning algorithms. There are several methods that have been proposed to extract features from such high dimensional data. This paper proposes a study of various methods for feature selection and it has been found that Clustering based Feature selection methods are most effective in selecting important features.**

Keywords: Feature selection, Search strategies, Machine learning, Clustering, Relevance.

I. INTRODUCTION

In recent years, social media services are used very widely emerging that allow people to communicate and express themselves conveniently and easily. The huge use of social media generates massive and high dimensional data. This poses new challenges to the task of data mining such as classification and clustering processes. One approach for handling such large scale and high dimensional data is Feature Selection.

Feature selection methods have been used form long years in the field of statistics and pattern recognition with the wide spread use of machine learning techniques [1]. Feature selection methods are needed when there are too much data that can be processed efficiently by machine learning algorithms, or when some features are costly to acquire and hence the minimum number of features are preferred [4].

Feature Selection methods are used to satisfy the common goal of maximizing the accuracy of classifier, reducing dimensionality, eliminating irrelevant and redundant data and improving result comprehensibility and helping to avoid slow execution time of learning algorithms [5].

II. RELATED WORK AND METHODOLOGY

The feature selection methods are divided into following categories according to their working principles [5].

1) methods which select the best subset of features that has a certain number of features

2) methods which select the best subset of features according to their own principles, independent of outside measures

A. Feature Selection Methods:

Feature Selection methods are also divided into several classes according to interaction with learning algorithms [2], [4], [8], [11].

1) *Filter method:* If feature selection method works independent from the learning algorithm, it is called a filter method. It consists of algorithms that are built in the adaptive systems for data analysis [11]. They use an evaluation function that relies on properties of data. These methods are fast, scalable and can be used with any learning algorithm effectively [4], [6], [8].

   Distance based and margin based criterion can be used for filters.

2) *Wrapper method:* If a feature selection method uses learning algorithm to guide its search process to weigh features, it is called a wrapper method. The algorithms of this method are wrapped around the adaptive systems providing them subsets of features and receiving their feedback [12]. These wrapper approaches are aimed at improving results of the specific predictors they work with [11]. It utilizes the classifiers as black box to find the subset of features based on their predictive power [15]. The wrapper methods are less scalable, may overfit the data more and are

ISSN: 2278 – 1323

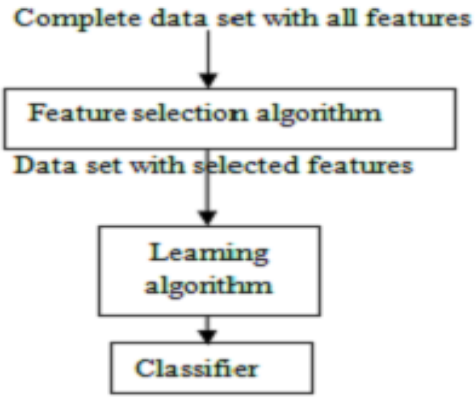

Fig. 1: Filter Method

more slower than filters because they require classifier training and validation. But the resultant features have more accuracy with specific learning algorithm [4], [6], [7], [8].

3) *Embedded method:* If a feature selection is embedded into a learning algorithm and optimized for it, it is called an embedded method. They are based on performance evaluation metric calculated directly from data, without direct reference to the results of any data analysis systems. These methods are faster than wrappers and make the most efficient selection for the learning algorithm that they collaborate with [4], [7].

Within the filter model, different feature selection algorithm can be further classified into two groups as feature weighting algorithms and subset search algorithms. This is based on whether they evaluate the goodness of features individually or through feature subsets [3].

1) *Feature weighting algorithms:* These assign weights to features individually and rank them based on their relevance to the target concept. A feature is good and thus will be selected if its weight of relevance is greater than a threshold value. The algorithm known as Relief is based on this criterion [21].

2) *Subset search algorithms:* These searches through candidate feature subsets guided by a certain evaluation measure which captures the goodness of each subset [22]. An optimal subset is selected when search stops.

Some evaluation measures, those remove irrelevant and redundant features, include the consistency measure [23] and the correlation measure [24], [25]. Consistency measure attempts to find a minimum number of features that separate classes as consistently as full set of features can where as an inconsistency is defined as two instances having the same feature values but different class labels.

In recent years, with respect to the filter feature selection methods, the application of cluster analysis is more effective than traditional feature selection algorithms [27], [28]. Distributional clustering of words reduces the dimensionality of text data. The graph-theoretic methods have been used in many applications those use cluster analysis. Graph theoretic clustering works like following: Compute a neighborhood graph of instances, and then delete any edge which is much larger or much shorter than its neighbors. The result we gain is forest and each tree in that forest is a cluster.

B. Feature Selection Process:

During feature selection process, all feature selection methods go through several phases. These are known as characteristic properties of feature selection method [6], [7].

There are several characteristic properties. These are Initial state of search, creating successors, search strategy, feature evaluation method used, including or not including the interdependence of features and halting criterion [6], [7].

1) *Initial State:* It is the condition of initial subset node in the search tree. I can be empty, full of features or can be filled with randomly selected features [6].

2) *Creating Successors:* It is about to making a forward or backward feature selection. The number of features that successors can have is

ISSN: 2278 – 1323

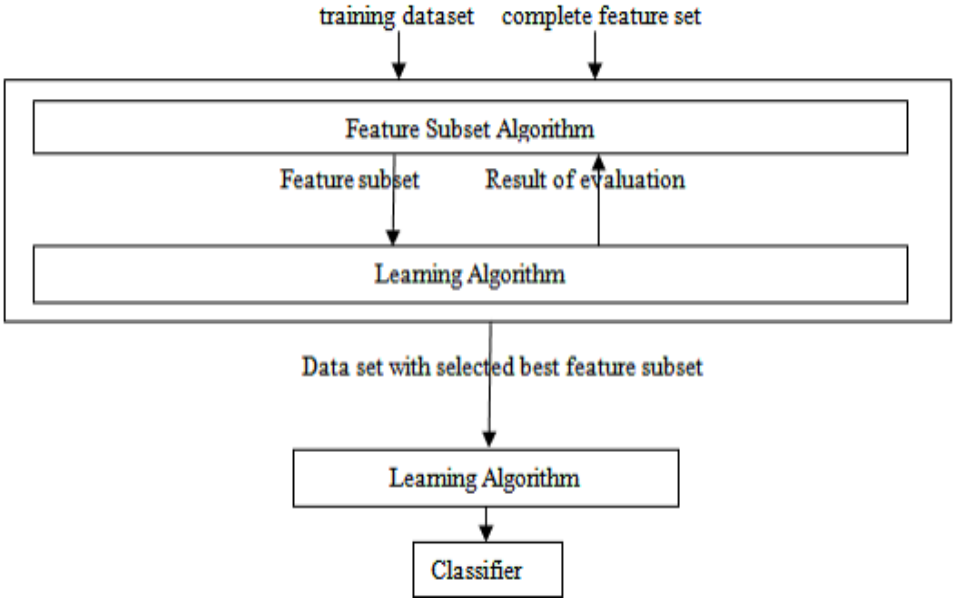

Fig. 2: Wrapper method

determined by this. Some compound methods combine both forward and backward methods [7].

3) *Search Strategy:* It is a strategy which is used to travel in the search tree. It can be exponential, sequential, or randomized [7], [9], [10].

4) *Feature Evaluation method:* These are used to give every feature a weight so that features can be compared to each other and efficient selection is possible [4], [8].

5) *Interdependence of features:* It determines whether the selection method is univariate or multivariate. Univariate methods only calculate the weight of features with their dependence to classes. Multivariate methods calculate the dependency between feature pair as well as their class relevance

6) *Halting criterion:* It is used to decide when to stop searching.

C. Search Strategy:

It is required to select candidate subset and objective function evaluates each of the candidate subset. There are different search strategies [16], [17].

1) *Exhaustive search:* It evaluates a number of subsets that grows exponentially with the dimensionality of search space. Each subset is evaluated by objective function and a measure of goodness is returned to the search algorithm [16], [17]. For this search, branch and bound algorithms are used. For example, Basic BB which is most slowest algorithm, Enhance BB which is most fastest algorithm, FAST BB and BB with partial prediction [18]. The time complexity is exponential in terms of dimensionality for exhaustive search [16], [17].

2) *Heuristic search:* Heuristics help to reduce the number of alternatives from an exponential number to a polynomial number [19]. For example, Sequential backward elimination, Sequential forward selection, and bidirectional search.

Forward Selection: It considers the subset to be empty initially and keeps on adding one feature at a time until the best feature subset is obtained.

Backward Elimination: It takes complete set of features as input and keeps on removing one attribute at a time until the most appropriate subset of features is obtained.

Heuristic search is problematic when the data consists of highly correlated features.

3) *Randomized search strategy:* It performs randomized exploration of search space where next direction is sample from a given probability [20]. For example, generic algorithm.

III. ALGORITHMS AND ANALYSIS

There are several algorithms based on above criterion to find out features related to work.

ReliefF algorithm relies on relevance evaluation [21]. The key idea of Relief is to estimate the relevance of features according to how well their values distinguish

ISSN: 2278 – 1323

between the instances of the same and different classes that are near to each other. Relief randomly samples a number (m) of instances from the training set and up-dates the relevance estimation of each feature based on the difference between the selected instance and the two nearest instances of the same and opposite classes. Time complexity of Relief is O(mMN) having data set with M instances and N features . However, Relief does not help with removing redundant features. As long as features are deemed relevant to the class concept, they will all be selected even though many of them are highly correlated to each other [21].

CFS algorithm exploits heuristic search. This subset search algorithm does not have strong scalability to deal with high dimensional data.

A novel algorithm named Fast Correlation-Based Filter Solution (FCBF) can effectively identify both irrelevant and redundant features with less time complexity than subset search algorithms [3].

The algorithm FCBF# has different search strategy than FCBF and it can produce more accurate classifiers for size k subset selection problem. It selects best subset of features from the full set by mean of backward elimination. This algorithm is good alternative for images and text data [13].

In recent years, Minimum Spanning Tree (MST) Clustering algorithms are most popular for feature selection, because they do not assume that data points are grouped around centers or separated by a regular geometric curve [26]. Recent algorithm named as Fast Clustering based Feature Selection Algorithm (FAST) is based on the MST method [14], [26]. This algorithm works like following:

1) Features are divided into clusters by using graph theoretic clustering methods.

2) The most representative feature which is strongly related to target class is selected from each cluster and final subset of features is formed.

As features in the different clusters are relatively different and independent, the clustering based FAST produces useful and independent features with high probability. FAST obtains rank 1 for Microarray data [26]. It is better than Relief-F.

IV. APPLICATIONS

Feature Selection is useful in many applications.

A. Text Classification:

Feature Selection has been applied to text categorization in order to improve its scalability, efficiency and accuracy. Since each document in the collection can belong to multiple categories, the classification problem is generally split into multiple binary classification problems with respect to each category. Accordingly, features are selected locally per category, for example, local feature selection [29]. There are number of feature selection metrics used for text categorization. Among those, Information Gain (IG), Chi-Square (CH), Correlation Coefficient (CC) and Odd Ratios (OR) are most effective [29].

B. Genre Classification:

Metadata such as filename, author, date, size, track length and genres are some common features used to classify and retrieve genre documents. By using this data, the classification is infeasible so the step to feature selection is required. For Genre classification, feature selection is a process where a segment of an audio is characterized into a compact numerical representation [30].

C. Microarray data analysis:

In most bioinformatics problems, the number of features is significantly larger than the number of samples [32]. For example, Breast Cancer Classification on the basis of microarray data, Network inference on the basis of microarray data. Thus this requires the feature selection. Content analysis and signal analysis in genomics also require feature selection.

D. Software defect prediction:

Two common aspects of data quality in software defect prediction that can affect classification performance are class imbalance and noisy attributes of data sets [31]. Feature selection methods like genetic algorithm and bagging technique can be used to improve the performance of the software defect prediction. Genetic algorithm is applied to employ with the feature selection, and bagging technique is applied to deal with the class imbalance problem [31].

E. Stock market analysis:

The analysis of the financial market always draws a lot of attention from researchers and investors. The trend of stock market is very complex. It is influenced by various factors. Therefore it is very important to find out the most significant factors to the stock market. Feature Selection is an algorithm that can remove the redundant and irrelevant factors, and finds the most significant subset of factors to build the analysis model [33].

F. Image Retrieval:

Feature selection can be applied to content based (color, shape, texture, etc) image retrieval. It is useful for efficient browsing, searching and retrieving of images.

ISSN: 2278 – 1323

V. CONCLUSION

This paper provides a brief overview of Feature Selection. Feature Selection is effectively used as a preprocessing step for various applications. Based on the review of existing literature, it may be appropriate to suggests that the clustering based feature selection methods are more efficient microarray data analysis. Also in that the minimum spanning tree clustering algorithm is most popular. The classification accuracy can be significantly improved with feature selection, where feature selection methods are applied with promising results.

ACKNOWLEDGEMENT

We would like to thank to the editors and the anonymous reviewers for their insightful and helpful comments and suggestions.

REFERENCES

[1] John G., Kohavi R., Pfleger K. "*Irrelevant features and the subset selection problem*", in Proceedings of the Eleventh International Machine Learning Conference (1994) 121–129. New Brunswick, NJ: Morgan Kaufmann.
[2] Guyon I., Elisseef A. "*An introduction to variable and feature selection*", J. Machine Learning Res. 3 (2003) 1157–1182.
[3] L. Yu and H. Liu, "*Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution*", In Proceedings of the Twentieth International Conference on Machine Leaning (ICML-03), 856-863, Washington, D.C., August 21-24, 2003.
[4] Saeys Y., Inza I., Larrañaga P. "*A review of feature selection techniques in bioinformatics*", Bioinformatics, 23(19) (2007), 2507–2517
[5] Liu H., Motoda H., "*Computational methods of feature selection*". Taylor, (2008).
[6] Blum A.L., Langley P., "*Selection of relevant features and examples in machine learning*", Artificial Intelligence (1997) 245-271
[7] Molina L.C., Belanche L., Nebot A., "*Feature Selection Algorithms: A Survey and Experimental Evaluation*". ICDM, (2002)
[8] Liu H., Yu L., "*Towards integrating feature selection algorithms for classification and clustering*". IEEE Transactions on Knowledge and Data Engineering, 17(3):1–12, 2005.
[9] Aha D., W. Bankert, R. L., "*A comparative evaluation of sequential feature selection algorithms*". In Doug Fisher and Hans-J. Lenz, editors, Learning from Data, chapter 4, pages 199-206. Springer, New York, 1996.
[10] Setiono R., and Liu H., "*A probabilistic approach to feature selection-a filter solution*". In Proceedings of International Conference on Machine Learning, 319-327 (1996).
[11] Wlodzislaw Duch, "*Filter methods*", Springer-Feature Extraction Studies in Fuzziness and Soft Computing Volume 207, 2006, pp 89-117.
[12] Ms. Shweta Srivastava, Ms. Nikita Joshi, Ms. Madhvi Gaur, "*A Review Paper on Feature Selection Methodologies and Their Applications*". International Journal of Engineering Research and Development e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.comVolume 7, Issue 6 (June 2013), PP. 57-61
[13] Baris Senliol, Gokhan Gulgezen, Lei Yu and Zehra Cataltepe, "*Fast Correlation Based Filter (FCBF) with a Different Search Strategy*". Computer and Information Sciences, 2008.ISCIS '08. 23rd International Symposium, 2008.
[14] A.GowriDurga, A.Gowri Priya, "*Feature Subset Selection Algorithm for High Dimensional Data using Fast Clustering Method*". IJCAT International Journal of Computing and Technology, Volume 1, Issue 2, March 2014 ISSN: 2348- 6090.
[15] L. Ladha et al., "*Feature Selection Methods and Algorithms*". International Journal on Computer Science and Engineering (IJCSE), Vol. 3. No. 5, pp. 1787-1797; 2011.
[16] Roberto Ruiz, Jos´e C. Riquelme, and Jes´us S. Aguilar-Ruiz; "*Heuristic Search over a Ranking for Feature Selection*"; IWANN, LNCS 3512, pp. 742–749; 2005.
[17] Yao-Hong Chan; "*Empirical comparison of forward and backward search strategies in L-GEM based feature selection with RBFNN*". International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 3, pp- 1524 - 1527 ; 2010.
[18] P. Somol, P. Pudil; "*Feature Selection Toolbox*". Pattern recognition, Published by Elsevier Science Ltd; 2002.
[19] Manoranjan Dash, Huan Liu; "*Consistency based feature selection*". Published by Elsevier
[20] Computer Science (Artificial Intelligence); 2003.
[21] Kira, K., & Rendell, L., "*The feature selection problem: Traditional methods and a new algorithm*". Proceedings of the Tenth National Conference on Artificial Intelligence (pp. 129–134). Menlo Park: AAAI Press/the MIT Press (1992).
[22] Liu, H., & Motoda, H. "*Feature selection for knowledge discovery and data mining*". Boston: Kluwer Academic Publishers (1998).
[23] Dash, M., Liu, H., & Motoda, H. (2000). "*Consistency based feature selection*". Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining (pp. 98–109). Springer-Verlag.
[24] Hall, M. (1999). "*Correlation based feature selection for machine learning*". Doctoral dissertation, University of Waikato, Dept. of Computer Science.
[25] Hall, M. (2000). "*Correlation-based feature selection for discrete and numeric class machine learning*". Pro-ceedings of the Seventeenth International Confer-ence on Machine Learning (pp. 359–366).
[26] Qinbao Song, Jingjie Ni and Guangtao Wang, "*A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data*", IEEE transactions on knowlwdge and data engineering VOL: 25 NO: 1 YEAR 2013
[27] 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.

[28] Dhillon I.S., Mallela S. and Kumar R., "*A divisive information theoretic feature clustering algorithm for text classification*", J. Mach. Learn. Res., 3, pp 1265-1287, 2003.

[29] Zhaohui Zheng, Xiaoyun Wu, Rohini Srihari, "*Feature Selection for Text Categorizationon Imbalanced Data*", Sigkdd Explorations, Volume 6, Issue 1 - Page 80
[30] Shyamala Doraisamy, Shahram Golzari, Noris Mohd. Norowi, Md. Nasir B Sulaiman, Nur Izura Udzir; "*A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music*" Proceedings of ISMIR, 2008.
[31] Romi Satria Wahono, Nanna Suryana Herman, "*Genetic Feature Selection for Software Defect Prediction*", Advanced Science Letters, Vol. 20, 239–244, 2014.
[32]Gianluca Bontempi, Benjamin Haibe-Kains; "*Feature selection methods for mining bioinformatics data*",

Kokane Vina A., PG Scholar, Department of Computer Engineering, JSPM's Bhivarabai Sawant Institute of Technology and Research, Wagholi, Pune.

Lomte Archana C., Assistant Professor, pursuing PHD, Department of Computer Engieering, JSPM's Bhivarabai Sawant Institute of Technology and Research, Wagholi, Pune.

ISSN: 2278 – 1323