Using Feature Selection Technique for Data Mining: A Review

Sanjograj Singh Ahuja[1], Srishti Arora[2]

[1,2]Student of Btech in Computer science, Manav rachna international institute of research and technology, Faridabad, India

---***---

Abstract - *In the pre-processing step, feature extraction is used to further develop the mining execution by decreasing information dimensionality. Despite various element determination calculations, it is still an active exploration region in information mining, AI, and example acknowledgement networks. Many component determination calculations defy extreme difficulties regarding adequacy and proficiency as a result of ongoing expansion in information dimensionality (information with a large number of elements or qualities or factors). This paper examines some current well-known element determination algorithms that address the rates and difficulties of those algorithms.*

***Key Words*: Feature selection, Classification, Data mining, Unsupervised learning, Filter strategy**

1.INTRODUCTION

In recent decades, information gathered for different examination objects is a lot bigger. Such informational index might comprise of thousands of examples (records) and every one of which might be addressed by hundreds or thousands of provisions (characteristics or factors) [1]. The high dimensional informational collection is the information that contains an amazingly enormous number of elements. DOROTHEA [2] is such a dataset utilized for drug revelation, comprises 1,950 occasions and 100,000 provisions. Many components in such informational collection contain valuable data for understanding the information applicable to the issue. Yet, it moreover includes a huge number of irrelevant elements and important excess stocks. This reduces the learning execution and computational accuracy [1]. To avoid this issue, a pre-processing step called "Element Selection" is utilized to decrease the dimensionality before applying any information mining procedures like Classification, affiliation rules, clustering and regression. The point of element choice is to decide an element subset as little as could be expected. It is the fundamental pre-processing venture preceding applying information mining assignments. It chooses the subset of unique elements with no deficiency of valuable data. It eliminates immaterial and repetitive elements for reducing information dimensionality. Accordingly, it further develops the mining accuracy, reduces the calculation time, and upgrades conceivability [3]. Applying mining undertakings to the decreased component subset creates a similar outcome with a unique high-dimensional dataset. Element determination offers benefits like lowering storage prerequisites, keeping away from overfitting, working with information representation, accelerating the execution of mining calculations and reducing preparing times [4]. This paper examines the methods utilized by an assortment of element choice algorithm, looks at their benefits and disservices, and assists with understanding the current difficulties and issues in this examination field.

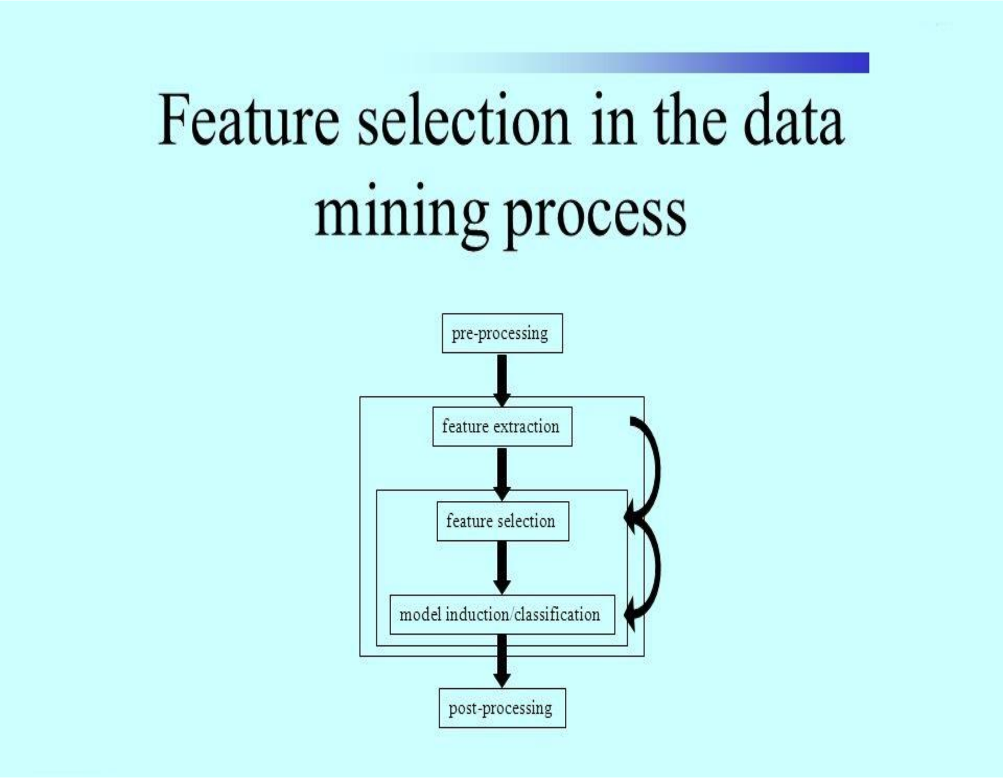
The rest of the paper is regulated as follows; in section 2, the essentials of component choice is examined. Existing element determination calculations are studied in section 3. section 4 finishes up our work.

2. PROCESS FEATURE SELECTION

The four key strides of a Feature selection interaction are highlight subset period, subset evaluation, reducing standard and result approval. The element subset time is a heuristic pursuit measure that brings about an up-and-comer subset for evaluation. It utilizes looking through techniques like total, consecutive and irregular pursuit to produce subsets of components. Dunne et al. [5] expressed that these studying through methodologies depend on stepwise expansion or cancellation of provisions.

The decency of the created subset is assessed utilizing an assessment rule. If the recently completed subset is superior to the past subset, it replaces it with the best subset. The last best element subset is then approved by earlier information or utilizing various tests. These two cycles are rehashed until the uncertain standard is reached. Fig.1 represents the

Table -1: Process of Feature Selection



Feature selection in the data mining process

pre-processing

feature extraction

feature selection

model induction/classification

post-processing

3. ALGORITHM OF FEATURE SELECTION

Based on the selection technique, including the selection of feature calculations are extensively grouped into three classifications: Filter, Wrapper and Hybrid Method [6]. Channel Method chooses the element subset based on inborn qualities of the information, unrestricted of mining algorithm. It tends to be applied to details with high dimensionality.

The benefits of the Filter strategy are its over-simplification and high estimation productivity. Covering Method requires a foreordained calculation to decide the best element subset. Prescient precision of the analysis is utilized for assessment. This strategy ensures better outcomes; however, it is computationally costly for the huge dataset. Hence, the Wrapper technique isn't typically favoured [7]. Mixture Method joins Filter and Wrapper to accomplish the benefits of both the methods. It utilizes an autonomous measure and a mining calculation to calculate the integrity of the recently produced subset [21]. In this methodology, the Filter technique is first applied to reduce the research space and afterwards, a covering model is used to acquire the best element subset [8]. Fig. 2 shows the crossover model.

Feature choice requires preparing information for learning purposes. The preparation information can be either named or unlabelled. According to the point of view of using mark data, feature determination calculations are characterized into supervised, unsupervised and semi-supervised measures [9]. Addressed include choice uses marked information for learning purposes while Unsupervised element determination utilizes unlabelled information.

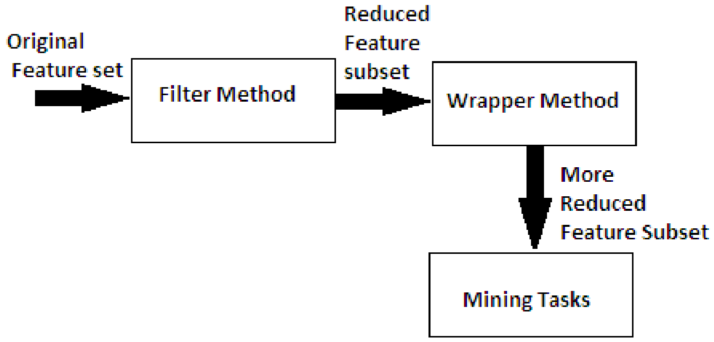


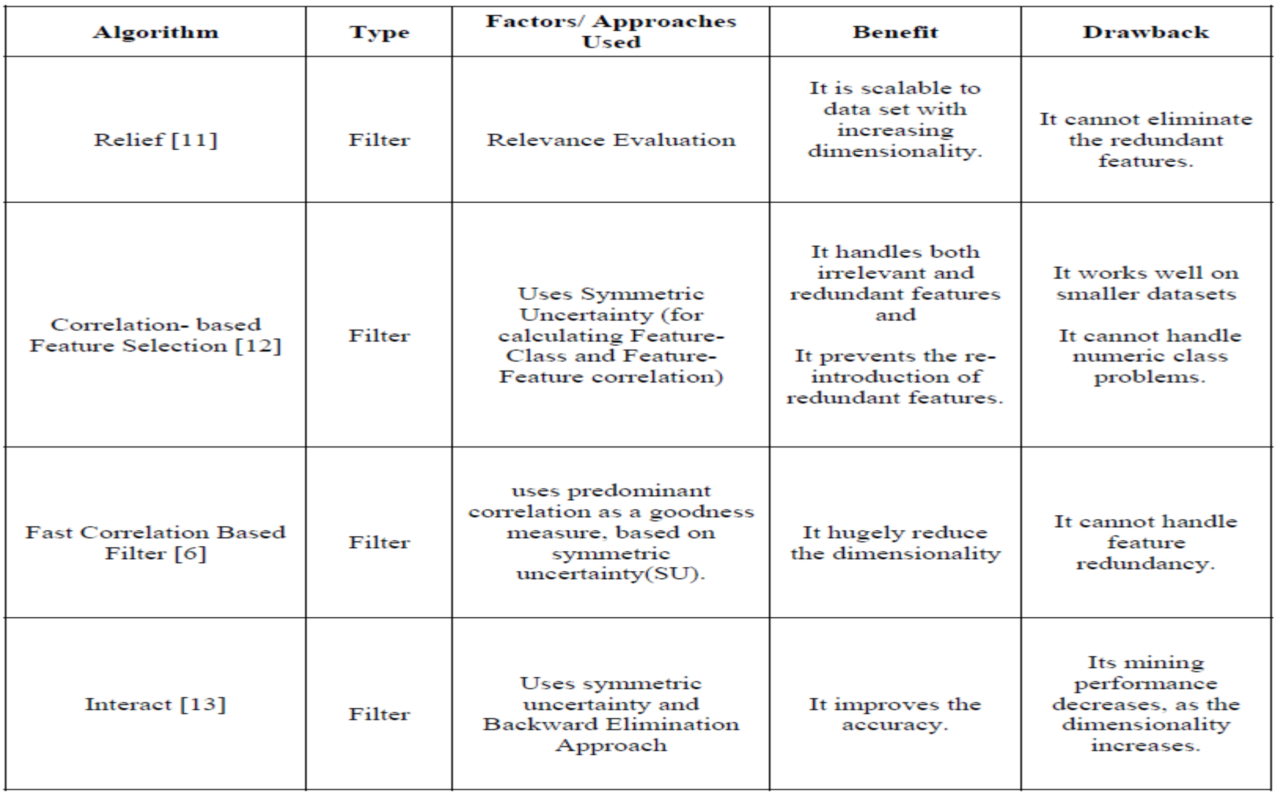
Fig 2: Hybrid Model

In managed learning, many named information is needed to accomplish better component determination execution [10]. Suppose the measure of detailed information is restricted. In that case, directed learning continues because the base measure of data is expected to guarantee that the connections between the objective ideas and the class names aren't accessible. This issue is alluded to as a "little marked example issue" [22]. Semi-supervised learning is another idea that expands the measure of named information by foreseeing the class marks of unlabelled information, which builds the learning execution.

4. FEATURE SELECTION ALGORITHMS COMPARISON

Feature Selection is the fundamental pre-processing step in Data mining. A few feature selection algorithms are accessible. Every algorithm has its solidarity and deficiency. Table 1 looks at a portion of the accessible measures.

TABLE I. COMPARISON OF SOME EXISTING FEATURE SELECTION ALGORITHMS



Algorithm	Type	Factors/ Approaches Used	Benefit	Drawback
Relief [11]	Filter	Relevance Evaluation	It is scalable to data set with increasing dimensionality.	It cannot eliminate the redundant features.
Correlation- based Feature Selection [12]	Filter	Uses Symmetric Uncertainty (for calculating Feature-Class and Feature-Feature correlation)	It handles both irrelevant and redundant features and It prevents the re-introduction of redundant features.	It works well on smaller datasets It cannot handle numeric class problems.
Fast Correlation Based Filter [6]	Filter	uses predominant correlation as a goodness measure, based on symmetric uncertainty(SU).	It hugely reduce the dimensionality	It cannot handle feature redundancy.
Interact [13]	Filter	Uses symmetric uncertainty and Backward Elimination Approach	It improves the accuracy.	Its mining performance decreases, as the dimensionality increases.

While choosing a component subset, the Filter strategy utilizes all the accessible preparing information; Filter strategies are quicker and better than coverings. It tends to be applied to huge datasets having many components [12]. However, Filter Method isn't in every case enough to acquire better precision [18]. Then again, Wrapper Method likewise chooses the best component subsets, yet it has been demonstrated to have high calculation costs when contrasted with Filter for huge datasets

[12]. Crossbreed technique is less computationally concentrated than covering strategies. Help [11] eliminates superfluous information utilizing the closest neighbour approach, yet It doesn't think about excess elements, though CFS and FCBF think about the repetitive provisions while choosing important aspects [8]. FCBF is a quick channel strategy. Quick calculation removes superfluous components just as it additionally handles excess elements [20]. It functions admirably with microarray information when compared and text and picture information [6]. Collaborate [13] and HFS [19] calculation further develops mining precision, yet it can't increase with the expanding dimensionality. CDMI [14] is noise-sensitive.

5. CONCLUSION

Among the current component choice algorithm, a few analyses include just in the choice of pertinent provisions disregarding excess. Dimensionality increments pointlessly due to excess components, and it additionally influences the learning execution. Furthermore, a few algorithms select significant parts, ignoring the presence of boisterous information. The presence of noisy information prompts helpless learning execution and builds the computational time. Our investigation infers a requirement for a successful structure for highlight determination, which should include in the best element subset with no repetitive and noisy information. It ought to be applied for a wide range of information, and it ought to likewise be ready to increase with increasing dimensionality.

REFERENCES

[1] Kashif Javed, Haroon A.Babri and Mehreen Saeed, "Feature Selection based on Class-Dependent Densities for High Dimensional Binary Data", IEEE Transactions on Knowledge and Data Engineering, Vol 24, No 3, 2012 (www.computer.org/csdl/trans/tk/2012/03/ttk2012030465-abs.html)
[2] "Feature Selection Challenge by Neural Information Processing Systems Conference (NIPS),"http://www.nipsfsc.ecs.soton.ac.uk,2003
[3] H.Liu and H.Motoda, Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, 1998.
[4] Zilin Zeng, Hongjun Zhang, Rui Zhang, Youliang Zhang, "Hybrid Feature Selection Method based on Rough Conditional Mutual Information and Naïve Bayesian Classifier", Hindawi Publishing Corporation, ISRN Applied Mathematics, Vol 2014, Article Id 382738,11 pages .
[5] K.Dunne, Cunningham and F.Azuaje, "Solution to instability problems with sequential wrapper-based approaches to feature selection", Journal Of Machine Learning Research,2002.
[6] Lei Yu, Huan Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", Department of Computer Science & Engineering, Arizone State University, Tempe, AZ 85287-5406, USA, 2003
[7] A.Blum and P.Langley, "Selection of relevant features and examples in machine learning", Artificial Intelligence, vol 97, pp 245-271, 1997
[8] Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data", IEEE Transactions on Knowledge and Data Engineering, Vol 25, No.1, 2013
[9] Z.Zhao, H.Liu, "On Similarity Preserving Feature Selection", IEEE Transactions on Knowledge and Data Engineering, Vol 25, no 3, 2013
[10] Yongkoo Han, Kisung Park and Young-koo Lee, "Confident Wrapper-type Semi-Supervised Feature Selection Using an Ensemble Classifier", IEEE, 2011
[11] K.Kira and L.A Rendell, "The Feature Selection Problem: Traditional methods and A New Algorithm," Proc. 10th National Conference Artificial Intelligence, pp.129-134, 1992.
[12] Mark A. Hall and Lloyd A. Smith, "Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper", Proceedings of the Twelfth International FLAIRS Conference, 1999.
[13] Zheng Zhao and Huan Liu "Searching for Interacting Features" Department of Computer Science and Engineering,Arizona State University, 2007
[14] Wang Liping, "Feature Selection Algorithm Based On Conditional Dynamic Mutual Information", International Journal O Smart Sensing and Intelligent Systems", VOL. 8, NO. 1, 2015
[15] Kexin Zhu and Jian Yang, "A Cluster-Based Sequential Feature Selection Algorihm", IEEE, 2013
[16] Y.Kim, W.Street, and F.Menczer, "Feature Selection for Unsupervised Learning Via Evolutionary Search," Proc. Sixth ACM SIGKDD International Conference, Knowledge Discovery and Data Mining, pp 365 – 369, 2000
[17] Hwang, Young-Sup, "Wrapper-based Feature Selection Using Support Vector Machine", . Department of Computer Science and Engineering, Sun Moon University, Asan, Sunmoonro 221-70, Korea, Life Science Journal 2014;11 (7)
[18] B.M Vidhyavathi, " A New Approach to Feature Selection for Data Mining", International Journal of Computational Intelligence Research, ISSN 0973-1873 Vol.7 Number 3, pp 263 – 269, 2011
[19] Jihong Liu, "A Hybrid Feature Selection Algorithm for Data sets of thousands of Variables" IEEE, 2010

[20] L.Yu and H.Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy", Machine Learning Research, Vol. 10, no. 5, pp 1205 -1224,2004

[21] Huan Liu and Lei Yu, " Towards Integrating Feature Selection Algorithms for Classification and Clustering", IEEE Transactions on Knowledge and Data Engineering, Vol.17 No.4 2005

[22] A.Jain and D.Zongker, "Feature Selection: Evaluation, Application and small sample performance", IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(2):153-158,1997