FASTER: A Hybrid Algorithm for Feature Selection and Record Reduction in Rare Frequent Itemset

Usman Qamar, Younus Javed

Abstract—The amount of data that has to be analysed and processed to assist decision making has significantly increased in recent years. These datasets may contain potentially useful, but as yet undiscovered, information and knowledge. This high dimensionality of datasets leads to the phenomenon known as the curse of dimensionality. When faced with difficulties resulting from the high dimension of a space, the ideal approach is to decrease this dimension, without losing relevant information in the data. The use of Rough-Set theory to achieve feature selection is one approach that has proven successful. However, most approaches carry out reduction in only one dimension i.e in the number of attributes. In this investigation a new algorithm is proposed which allows for record reduction as well as attribute reduction. FASTER (FeAture SelecTion using Entropy and Rough sets) is a hybrid pre-processor algorithm which utilizes entropy and rough-sets to carry out record reduction and feature (attribute) selection respectively. FASTER produced an attribute reduction of 30% with a speed improvement of 2.6 times when used as pre-processor for two different rare itemset algorithms.

Keywords- Rough-sets, Classification, Feature Selection, Entropy, Outliers, Rare itemset mining.

I. Introduction

The task of feature selection is to select a subset of the original features present in a given dataset that provides most of the useful information [1]. Hence, after the selection process has taken place, the dataset should still have most of the important information still present in it. In fact, a good FS techniques should be able to detect and ignore noisy and misleading features. The most intuitive method for feature selection is to enumerate all the candidate subsets. Unfortunately, exhaustive search is infeasible in most circumstances as there are 2n subsets for a feature set of size n. Hence, exhaustive search can only be used in domains where n is relatively small; a large n will make the search intractable in many real world applications. An alternative way is to use a random search method where the candidate feature subset is generated randomly [2]. Each time, the evaluation measure is applied to the generated feature subset to check whether it satisfies certain criteria. This process repeats until one subset that satisfies the given criteria is found. The process may also end when a predefined time period has elapsed or a predefined number of subsets have been tested. The third and most commonly used method is called heuristic search [2], where a heuristic function is employed to guide the search. The search is directed to maximize the value of a heuristic function.

As already mentioned the aim of feature selection is to remove unnecessary features from a set of attributes. Unnecessary features can be classified as irrelevant features and redundant features [1]. Irrelevant features are those that do not affect the target concept in any way, whilst redundant features do not add anything new to the target concept. These are two feature qualities that must be considered by FS methods i.e. relevancy and redundancy. An informative feature is one that is highly correlated with the decision concept(s) but is highly uncorrelated with other features. Similarly, subsets of features should exhibit these properties of relevancy and non-redundancy if they are to be useful.

In [3] two notions of feature relevance, strong and weak relevance, were defined. If a feature is strongly relevant, this implies that it cannot be removed from the dataset without resulting in a loss of predictive accuracy. If it is weakly relevant, then the feature may sometimes contribute to accuracy, though this depends on which other features are considered. These definitions are independent of the specific learning algorithm used.

Manuscript received Feb 18, 2014.
Usman Qamar and Younus Javed are with the Computer Engineering Department, College of Electrical and Mechanical Engineering, National University of Sciences and Technology. Islamabad, Pakistan (e-mail: usmanq@ceme.nut.edu.pk, myjaved@ceme.nust.edu.pk).

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Rough-Set Theory (RST) [4] is a pre-processing method to reduce dataset dimensionality before some other action is performed (for example, classification, clustering etc).

RST was proposed by Pawlak for knowledge discovery in datasets [4]. Certain attributes in an information system may be redundant and can be eliminated without losing essential information. Given a dataset with discretized attribute values, it is possible to find a subset of the original attributes using RST that are the most informative: all other attributes can be removed from the dataset with minimum information loss. Unlike statistical correlation-reducing approaches, it requires no human input or intervention. Most importantly, it also retains the semantics of the data, which makes the resulting models more transparent to human scrutiny.

The use of RST to achieve feature selection is one approach that has proven successful. Over the past twenty years, RST has become a topic of great interest to researchers and has been applied to many domains (e.g. classification [5, 6, and 7], systems monitoring [8], and data clustering [9]).

The curse of dimensionality is not limited only to attributes, and approaches can be extended to reduce the number of records in a dataset. As an example, let's consider itemset mining i.e the process of determining which groups of items appear together. Itemset mining can be divided into two main categories: Frequent Itemset Mining and Rare Itemset Mining.

1. Frequent itemset mining: This type of itemset mining is focused on determining which groups of items frequently appear together in transactions.
2. Rare Itemset mining: In some situations it may be interesting to search for "rare" itemsets, i.e. itemsets that do not occur frequently in the data (contrasting frequent itemsets). These correspond to unexpected phenomena, possibly contradicting beliefs in the domain.

One technique that can be used to reduce the number of samples or records in a dataset is entropy. Entropy is derived from information theory [10], which analyzes the information content of a dataset using different information theoretic measures. The idea behind such techniques is that normal data is regular in terms of a certain information theoretic measure. On the other hand, outliers may significantly alter the information content of the data because of their surprising nature (or they may be noise). Thus, these approaches detect data instances that induce an irregularity in the data, where the regularity is measured using a particular information theoretic measure such as entropy. Outlier detection using entropy is based on the observation that 'removing outliers from a dataset will result in a dataset that is less dissimilar' [10].

The aim of this investigation is to propose, develop and investigate a new algorithm which allows for both record reduction and attribute reduction in the domain of itemset mining.

II. ROUGH SET THEORY

RST determines the degree of attributes dependency and their significance.

An information system (IS) ([4]) is basically a flat table or view. An IS (Λ) is defined by a pair (U,A), where U is a non-empty, finite set of objects and A is a non-empty, finite set of attributes [10].

$$\Lambda = (U, A)$$

Every attribute $a \in A$ of an object has a value. An attribute's value must be a member of Va which is called the value set of attribute a [4].

$$a : U \rightarrow V_a$$

Decision systems (DS) [4] are a special kind of IS. By labeling the objects of A, it is possible to construct classes of objects. These classes can then be modeled using rough set

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

analysis. The labels are the target attribute of which to obtain knowledge.

A decision system (i.e. a decision table) expresses all the knowledge about the model. This table may be unnecessarily large, in part because it is redundant in at least two ways: the same or indiscernible objects may be represented several times, or some of the attributes may be superfluous.

In practice most sets cannot be determined unambiguously and hence have to be approximated [4]. This is the basic idea of rough sets. If IS A = (U,A) and $B \subseteq A$ then it is possible to approximate decision class X using the information contained by the attribute set of B. The lower and upper approximations are defined as follows [3]:

$$X : \underline{B}X = \left\{ x \middle| [x]_B \subseteq X \right\}$$

$$X : \overline{B}X = \left\{ x \middle| [x]_B \cap X \neq 0 \right\}$$

The lower approximation [4] contains all objects that are certainly members of X. The objects in the set of the upper approximation [4] are possible members of X. The boundary region [4] is defined as the difference between the upper and the lower approximation.

$$X : BN_B(X) = \overline{B}X - \underline{B}X$$

One natural dimension of reducing data is to identify classes, i.e. objects that are indiscernible using the available attributes. Savings are to be made as only one element of the equivalence class is needed to represent the entire class. The other dimension of reduction is to keep only those attributes that preserve the indiscernibility relation and, consequently, set approximation. The remaining attributes are redundant as their removal should not worsen the classification. There are usually several such subsets of attributes and those which are minimal are called reducts. Computing equivalence classes is straightforward; in contrast, finding the minimal reduct is NP hard [4].

Computing reducts is a non-trivial task that cannot be solved by a simple increase of computational resources. It is, in fact, one of the bottlenecks of the rough set methodology [4]. Fortunately, there exist good heuristics based on genetic algorithms that compute sufficiently many reducts in often acceptable time [4].

III. ENTROPY BASED OUTLIER DETECTION

A popular method for generating outliers or rare records using entropy is the local-search heuristic based algorithm (LSA) [10]. An LSA detects rare records based on the concept of minimizing the entropy of the dataset.

Suppose that, for a dataset D, k outliers are to be detected using entropy. The value of k is defined by the user. Initially, the set of outliers (denoted by OS) is specified to be empty and all the dataset's records are marked as non-outliers. k scans over the dataset are carried out to select k records as outliers In each scan, each record labelled as a non-outlier is temporarily removed from the dataset and the change in entropy is evaluated. The record that achieves the max entropy change, i.e., the maximum decrease in entropy experienced by removing that record, is selected as the outlier in each current scan and added to OS. This continues for each scan until the size of OS reaches k.

IV. IMPLEMENTATION

The FASTER algorithm can be divided into three main phases. The first phase uses entropy to generate an initial set of outliers and grade each record in the dataset in terms of its likeliness or unlikeliness of being a rare record; the second phase carries out feature selection and attribute reduction using rough-sets; while the third phase of the algorithm utilizes the grading of records to remove any redundant records.

1. Phase 1

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

The first phase of the algorithm uses entropy to generate an initial set of outliers. These outliers are not the final set but are used as decision attribute by rough-sets for the purpose of feature selection and attribute reduction in Phases and 3. In addition to this, the change in entropy value for each record is used to grade the record in terms of its likeliness or unlikeliness of being a rare record.

2. Phase 2

The second phase of the algorithm carries out feature selection using rough-sets. Feature selection removes redundant and insignificant attributes, thus generating a smaller, focused set of attributes.

3. Phase 3

Phase 3 of FASTER uses the grading score of each record in terms of its likeliness or unlikeliness of being a rare record. Using this information any records which are found to be least significant and thus whose removal should not affect the data are then removed.

Although Phases 2 and 3 can be swapped, this will have a negative effect on the quality of the reducts that are generated by rough-sets. This is because each record in the dataset is given a decision attribute in phase 1 i.e. outlier or non-outlier by OutlierAlg. This decision attribute is used by rough-sets to carry out feature selection. However if the record is removed i.e. Phase 3 is carried out before Phase 2, the decision attribute may itself be removed. As rough-sets use decision attributes for the purpose of generating reducts, the removal of the decision attribute will have a direct impact on the accuracy of the reducts. The quality of reducts generated by rough-sets is critical for FASTER: the better the quality of reducts, the higher the accuracy of FASTER.

In summary, the phases of developing the FASTER are:

- Phase 1: Generating outliers using an Entropy Outlier Algorithm
- Phase 2: Generating Reduct Attributes using Rough-Sets
- Phase 3: Removing Records

The main concept of FASTER is to divide the dataset into small portions and then process these small portions several times instead of using the entire dataset. This should enable the generation of a more consistent set of reducts. This process is very similar to generating dynamic reducts. Those reducts frequently occurring in random subtables can be considered to be stable and consistent. For this experimentation the reducts that appear more than once i.e. twice or more are selected as consistent reducts.

V. EXPERMINTATION

FASTER was applied as a pre-processor to seven datasets. Each dataset has a different number of attributes varying from 24 to 60. The size (number of records) of the datasets also varied from 6000 to 250,000 as well as the composition of numerical and categorical attributes.

Table I: Dataset Descriptions For Rare Itemset Mining

Dataset No	Dataset Name	No of Categorical Attributes	No of Numeric Attributes	Source
D1	SARS 2001 (a) census data	10	12	[11]
D2	SARS 2001 (b) census data	12	15	[11]
D3	SARS 1992 (a) census data	16	11	[11]
D4	SARS 1992 (b) census data	14	19	[11]
D5	Wisconsin breast cancer (a)	7	7	[12]
D6	Wisconsin breast cancer (b)	18	10	[12]
D7	HSV patients	10	14	[12]

FASTER is applied as a pre-processor for two different rare itemset mining algorithms i.e. 1) SUDA 2) MINIT.

SUDA [13] was motivated by statistical disclosure control (SDC). If confidential information is released, even in an anonymised form, there is a risk of individuals being identified using statistical disclosure through the matching of

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

known information with the anonymised data and disclosure of material specific to those individuals [13]. This leads data providers to apply SDC techniques to the data, variously recoding, masking and perturbing the data in order to reduce the statistical disclosure risk. An important aspect of disclosure control is the identification of 'risky' or special unique records, i.e. records whose unique status arise from having an unusual combination of a small number of attributes (e.g. a 16-year-old widow) [13]. Special uniques can be distinguished from random uniques, which are unique merely by the way key attributes are coded and therefore whose unique status is sensitive to variations in coding [13]. The ability to locate and grade all special unique records within a dataset enables more efficient disclosure control and improves the quality of released data.

SUDA has been designed specifically for this problem. It has been developed for discrete data (both numerical attributes and numerically coded categorical data) and can accept continuous data if it is transformed into a discrete form beforehand (via multiplication by factors of 10 and/or rounding up). SUDA can then find risky records or outliers before the datasets are released.

Minimal Infrequent Itemsets (MINIT) [14] is an infrequent itemset mining algorithm which falls under the category of rare itemset mining. MINIT can be used to find infrequent itemsets in statistical disclosure risk assessment, bioinformatics, and fraud detection. The two algorithms differ in terms of input datasets [14]. The easiest way to describe the differences in dataset properties is to consider the matrix form. For traditional itemset mining, the matrix consists of binary entries [14]. But for SUDA, the matrix entries can contain any integer [14]. We can transform a SUDA-type matrix into a binary matrix by enumerating all of the <column, value> pairs [14]. For each of these pairs, a column is created in the transformed binary matrix. For every value in a column in the SUDA-type input matrix, the corresponding <column, value> location in the transformed binary matrix is given a one. MINIT has been designed to handle the more traditional dataset definition. Details of the working of MINIT can be found in [14].

Table II: FASTER Improvement Results

	Average % Attribute Reduction	Average % Records Removed	Average Speed-up Achieved
SUDA	30%	12%	2.2x
MINIT	40%	18%	2.6x

FASTER was able to reduce the number of attributes on average by 30% for SUDA. It was able to reduce the number of records by 12% on average for SUDA and finally it was able to achieve an average computation speed of 2.2 times when compared with original algorithm. Similarly for MINIT, FASTER was able to reduce the number of attributes on average by 40%. It was able to reduce the number of records by 18% on average and finally it was able to achieve an average computation speed of 2.6 times when compared with original algorithm. These results clearly demonstrate that FASTER is a suitable candidate as a pre-processor for rare itemset mining.

VI. CONCLUSION

FASTER is a hybrid pre-processor algorithm which utilizes both entropy and rough-sets to carry out feature selection. The aim of FASTER is to reduce dimensions both horizontally and vertically i.e. columns corresponding to attributes and rows corresponding to number of distinct samples or records. FASTER can be divided into three main phases. The first phase uses entropy to generate an initial set of outliers and grade each record in the dataset in terms of its likeliness or unlikeliness of being a rare record. The second phase carries out feature selection and attribute reduction using rough-sets, while the third phase of the algorithm utilizes the grading of records to remove any redundant records. Results clearly demonstrate that FASTER is a suitable candidate as a pre-processor for rare itemset mining.

REFERENCES

[1] R. Jensen, Q. Shen. Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches. IEEE Press/Wiley & Sons, September 2008.

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[2] J. Huaa, W. D. Tembeb, E. R. Dougherty. Performance of feature-selection methods in the classification of high-dimension data, Pattern Recognition Volume 42, Issue 3, pp 409-424, 2009.
[3] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. Proceedings of 11th International Conference on Machine Learning. Morgan Kaufmann, pp. 121–129. 1994.
[4] Pawlak. Rough Sets: Theoretical Aspects of Reasoning About Data. Dordrecht: Kluwer Academic. 1991.
[5] Li-Juan, L. Zhou-Jun, A novel rough set approach for classification, IEEE International Conference on Granular Computing, pp. 349- 352, 2006.
[6] C. Hung, H. Purnawan, B,Kuo, Multispectral image classification using rough set theory and the comparison with parallelepiped classifier, Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International, pp. 2052-2055, 2007
[7] R. Jensen and Q. Shen. Fuzzy-rough data reduction with ant colony optimization. Fuzzy Sets Systems, vol. 149, Issue No. 1, pp. 5–20. 2005.
[8] S. Zhao, C. C. Tsang, On fuzzy approximation operators in attribute reduction with fuzzy rough sets, Information Sciences: an International Journal archive, Vol. 178, Issue 16, pp. 3163-3176, 2008.
[9] C. Bean, and C. Kambhampati, Autonomous clustering using rough set theory. International Journal of Automation and Computing , Vol.5 (No.1). pp. 90-102. 2008
[10] J. Liang, Y. Qian, Information granules and entropy theory in information systems, Science in China Series F: Information Sciences, Vol. 51, pp. 1427-1444, 2008.
[11] Samples of Anonymised Records (SARs), http://www.ccsr.ac.uk/sars/
[12] UCL Machine Leraning Group.
[13] M Elliot ., Manning A.M., and Ford R.W.: A computational algorithm for handling the special uniques problem. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems. Vol. 10, pp. 493-509, 2002.
[14] David J. Haglin, Anna M. Manning: On Minimal Infrequent Itemset Mining. DMIN, pp141-147, 2007.

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)