International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 3 (2015) pp. 6817-6826
© Research India Publications
http://www.ripublication.com

Infrequent Weighted Item Set Mining Using Decision Making Approach Algorithm

A.Mary Posonia [1] and J.Kanmani Rajathi [2]

[1] Asst. Professor, Faculty of Computing, Sathyabama University Chennai, Tamil Nadu, India
[2] Student, Department of Computer Science, Sathyabama University, Chennai, Tamil Nadu, India
Email: kanmani.j.cse @gmail.com

Abstract

Data mining process is used to analyze data from different perspectives and summarize it into necessary information. It can be used to extract information from datasets and transform it into an understandable structure for future use. Each data is to be represented by weights in order to extract the data quickly. An item set consists of collection of data's that are given weights based on data features. Weights are assumed to the item set by the frequent and infrequent use of the item set. To catch an infrequent weighted item set, is very intricate. To overcome the exception of infrequent weighted item set some algorithms and new methods shall be used. This method will excerpt the infrequent weighted item set by assuming threshold value based on transactions and finding minimal infrequent weighted item sets using SVM classifier. Finally decision making approach is made to remove the infrequent Item sets from the list in order to minimize the cost function.

Key Terms— Distributed Function, Decision making approach, Item set mining and SVM classifier.

1. INTRODUCTION

Item set mining is an exploratory data mining technique widely used for discovering valuable correlation among records. Numerous attempts to attain item set mining was focused on discovering frequent item sets, as a result the patterns are observed as the frequency of occurrence in the source data (the support) is above a given threshold. Persistent item sets finds application in a number of real-life contexts such as market basket analysis, medical image processing and biological data analysis. However, many traditional approaches ignore the influence or interest of each item or

transaction within the analyzed data. To permit treating items or transactions differently based on the significance in the frequent item set mining procedure, the proposal of weighted item set has been presented. A weight is associated with each data item and characterizes its local significance within each transaction. This method will address the discovery of infrequent and weighted item sets, from transactional weighted data sets.

2. RELATED WORKS

2.1 Mining

According to [2], Mining is an interdisciplinary subfield of computer science, is the computational progression of examining forms in hefty data sets containing methods at the lump of artificial intelligence, machine learning, statistics, plus database systems. The entire aim of the data mining process is to mine evidence from a data set and renovate it into a reasonable structure aimed at supplemental custom. Apart from the rare investigation step, it involves databank plus data management types, data pre-processing model in addition inference discussions, interestingness metrics, complexity deliberations, post-processing of uncovered structure, revelation and online updating. A records set is a group of data. Most usually a data set parallels to the matters of a discrete record counter, or a retiring arithmetical data matrix, where each column of the table indicates a specific variable, and each row parallels to a given member of the data set in demand. The data set list standards for each variables such as altitude and load of an object, for each associate of the data set. Each value is called as a datum. The data set may hold data for individual or added members, relating to the sum of rows.

2.2 Support Vector Machine

Support vector machines [3] continues supervised learning models by means of associated learning algorithms that examine facts and identify outlines, secondhand for classification and regression scrutiny. Assume a standard set of training samples, each marked as fitting to unique of dual clusters, an SVM [3] training algorithm figures a classic that allots fresh illustrations into one group or an additional group, crafting it as a non-probabilistic dual linear classifier. An SVM [3] model is a description of the patterns as ideas in space, plotted so that the instances of the distinct kinds are separated by a clear gap that is as extensive as probable. Fresh instances are then mapped into that same space and anticipated to belong to a category based on which side of the gap they fall on. In accumulation to execution of linear classification, SVM's can proficiently achieve a non-linear classification consuming what is named the kernel trick, obliquely mapping their inputs hooked on high-dimensional feature spaces.

2.3 Minimal Infrequent Item set Mining

D.J. Haglin and A.M. Manning stated that [4], the significant of discovering association rules is to set additional to the tricky of ruling item set that seem recurrently in a dataset. Applications of ruling infrequent item set contain statistical

disclosure risk assessment, bioinformatics, plus fake recognition. Fewer considerations has been rewarded to infrequent item set. The Infrequent item set recycled in several potential applications embraces statistical disclosure, bioinformatics and fake recognition. A pioneering algorithm can be enhanced to switch to the additional traditional dataset explanation and to hold finding minimal τ - infrequent item set MIIs. This can be named as MINIT, for **MIN**imal **I**nfrequent i**T**emsets. Primarily, a standing of objects is arranged by computing the provision of each of the items and then crafting a tilt of items in arising law of support. Minimal τ –infrequent item sets are exposed by allowing for every item ij in rank demand, recursively calling MINIT on the support set of the dataset with respect to ij considering only those items with sophisticated rank than ij, and then to prove each candidate Mij besides the original dataset. To the above delinquent innovative algorithm MINIT algorithm is used to discovery minimal τ -infrequent or minimal τ-occurrent item sets. The computation time obligatory on the datasets proposes a correlation among the numeral of MIIs and the aggregate of computation time required. It would be fascinating to see how fine MINIT could run in a parallel or grid environment. It would similarly be beneficial to discover supplementary pruning schemes to progress the running time necessities.

2.4 Association Rule Mining without Pre-assigned Weights

K.Sun and F.Bai acknowledged that [5], the association rule mining is a vital dispute in data mining. The usual facsimile ignores that the variance are flanked by the relations, and the slanted association rule mining does not work on databases with merely binary attributes. The elucidation to the delinquent can be as diverse procedures of w-support, which does not engage the preassigned tons. It precedes the characteristics of the dealing into consideration expending the link-based models. A wild mining algorithm is given, an enormous amount of investigational outcomes are existing. A comprehensive description of HITS is practical to the chart to rank the particles, wherever all nodes and links are endorsed to have weights. However, the model has a drawback that it only ranks items but does not deliver a measure like weighted support to estimate the random item set. It may be the key prolific try to smear link-based models to association rule mining. Therefore, the investigational upshot shows that the computational cost of the link-based model is realistic. At the outflow of three or four supplementary database scans the learned fallouts different from those acquired by traditional counting-based models. Mostly for auxiliary data sets, several significant item sets that are not so recurrent can be originated in the link centered model.

2.5 Efficient Pattern Mining

C.K.S. Leung, C.L. Carmichael, and B. Hao examined that [6], mining the frequent item set from transactional datasets exhausting algorithm is a good elucidations. In the incident of unclear data, however, numerous innovative techniques have been considered. Unfortunately, the suggestions often suffer when a lot of items results with the various changed probabilities. In the frequent set mining, the transaction dataset is usually signified as a binary matrix where every line characterizes a

transaction and each column relates to an item. An element Mij denotes the presence or the absence of the item j in transaction i by the value 1 or 0 respectively. For this the simple outdated model, where an item is either present or absent in a transaction several algorithms have been recommended for mining frequent item set. A new suggested method entitled as U-Eclat based on sampling by instantiating "possible worlds" of the undefined data, on which the algorithm subsequently runs boosted frequent item set mining algorithms. This will increase ability at a extremely little loss in accuracy. This is established for the statistical and an pragmatic evaluation on real and synthetic data. The only variation for U-Eclat comprises in impression the undefined contacts and instantiating. The U-Eclat algorithms will instantiate the uncertain datasets from numerous experimentations. The higher the numeral instantiations, the improved is the exactness of the outcome developed, at the outlay of an growth in accomplishment time. The experimented U-Eclat algorithm is associated with the innovative ECLAT and FP growth algorithms. It produces enhanced outcomes than the supplementary algorithms. Obviously the scope of these datasets develops very huge for numerous repetitions, and thus, those experiments always ensued in shrinkage in performance as paralleled to other algorithms.

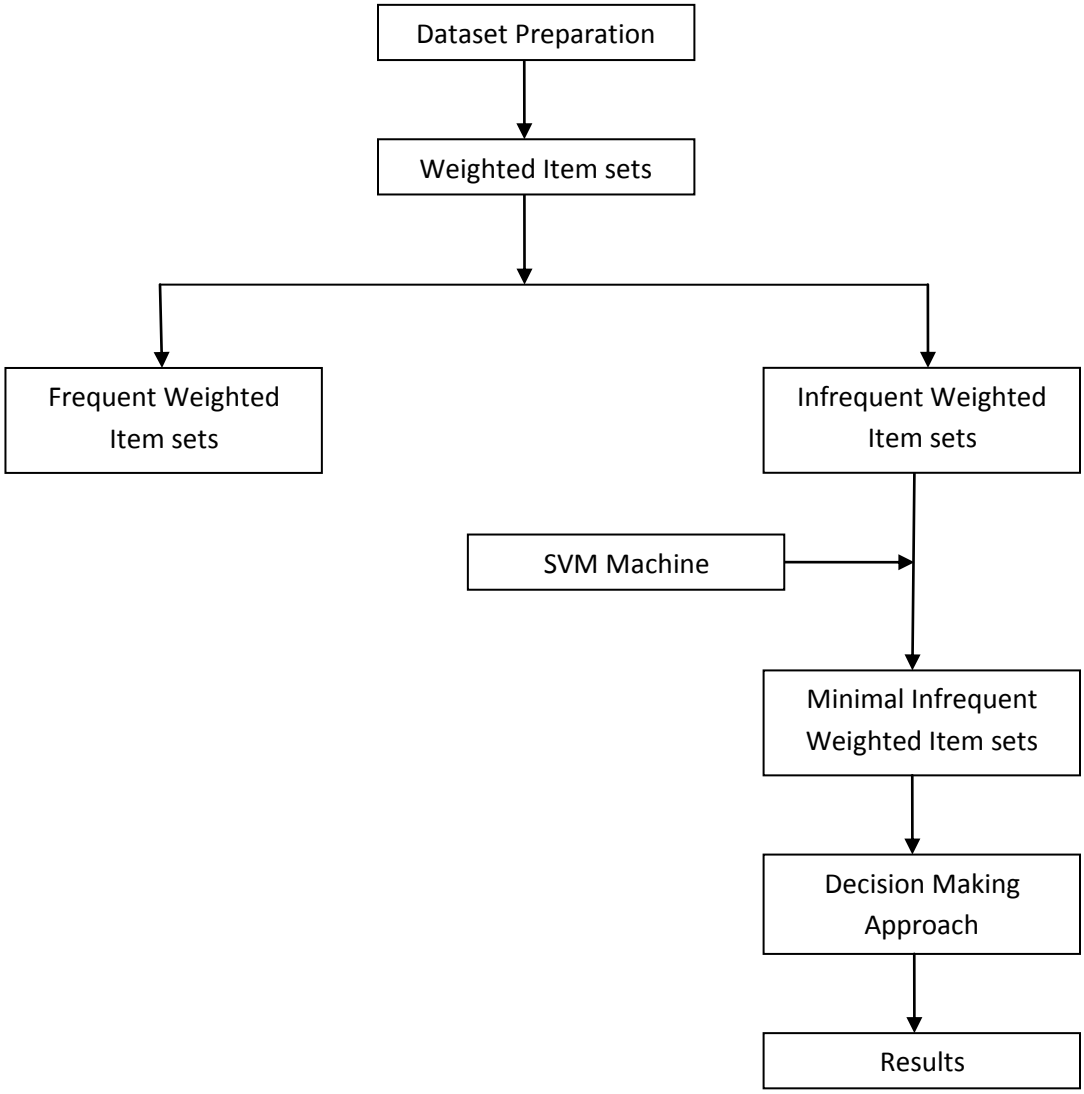
3. PROBLEM STATEMENT

The key chore of the development is

- To discover the Infrequent weighted item set by supposing threshold value.
- To treasure the minimal infrequent weighted item sets using SVM Classifier.
- To device a decision based tactic to trifle the infrequent weighted item sets.
- To lessen the cost function.

4. PROPOSED SYSTEM

The planned system outfits Support Vector Model and a decision based method. The proposed method ponders a real-time item set and accepts weight value to all the item sets. Based on the transactions the weight value will be improved or reduced. A threshold value is static to discover the frequent and infrequent item sets. The minimal infrequent item sets is found-out by SVM [3] classifier. SVM [3] classifier comprises of two phases namely training phase and testing phase. Primarily a model data is taken and skilled in the SVM [3] classifier using probability distribution function. The frequent item sets are smeared into the SVM [3] testing phase. The testing phase will trial the input data with the trained data using probability distribution function. The categorized data is deliberated as the minimal infrequent item set. To conclude a decision based approach is projected to eliminate the minimal infrequent item sets from the database in order to lessen the cost function. This technique uses SVM [3] classifier to find the minimal infrequent weighted item sets. The architecture diagram is given as below.



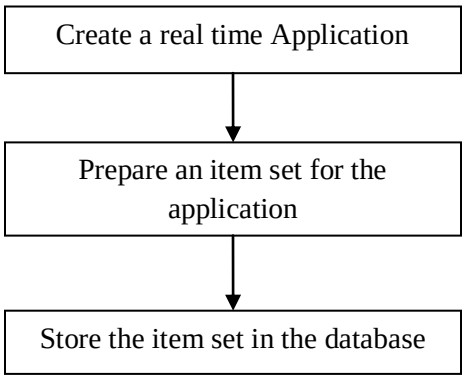
After ruling the minimal infrequent weighted item set, a decision making approach is prepared to eliminate the minimal infrequent item sets from the list in order to abate the cost function. The benefits of the method are it is a calm and simple method, diminishes the time complexity and cost function. An algorithm is advanced based on the SVM [3] classification. This algorithm will treasure the minimal infrequent weighted item sets by training and testing methods. Primarily, a trial minimal infrequent item sets are trained in the SVM [3] training phase after that when new item sets enters the testing phase it will contest the input item sets with the trained item sets using probability distribution function.

The system can be as, first the datasets are collected. The dataset can be a real time application or online application. Then the data are being weighted and categorized into frequent item sets and infrequent items sets based on the threshold value. The frequent item set is set as such and no further classification is being

processed. The infrequent item set is identified using support vector machine, the set of data is being fed into the machine and trained. The minimal infrequent item set is poised and it is trained using the machine. The decision making approach algorithm is used to remove the most infrequent data and thus this method reduces the cost.

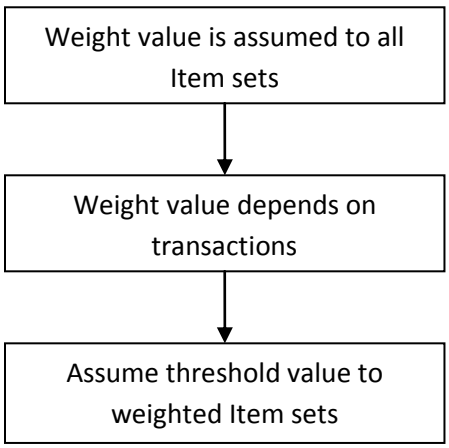
DATASET PREPARATION

The item set is prepared based on some of the real time applications. This item set comprises of records and attributes of the application and includes the features such as its works, operations performed and study of records are all warehoused in the database.



The data sets of any online application or real time application are collected. In this segment, the data of a real time application is constructed and the item for the particular application is prepared. When the item sets are organized it is once stored in a particular database.

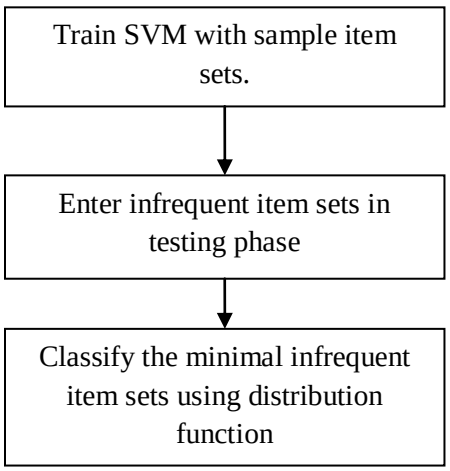
FINDING INFREQUENT WEIGHTED ITEMSETS



In this phase, the item sets are assumed to the weights and depends on the transactions of the item sets. The weight value rises based on the transaction of the item sets. A threshold value is assumed to each of the weighted item sets. The item sets lesser than the threshold value are reflected as infrequent weighted item sets. There are three phases in this module. The first step is each of the item sets is given a specific weight. The weight is assigned to each of the item sets based on the number of times the transaction has been occurred. The threshold value is set to each of the item set and the ordering of the item sets depends only on the threshold value that is being assigned. The item sets which are higher than the threshold value is deliberated to be as the frequent weighted item set i.e. most frequently used item sets.

IDENTIFY MINIMAL INFREQUENT WEIGHTED ITEMSETS

In this stage some sample minimal infrequent item sets are trained with SVM [3] classifier. Once the infrequent item sets are generated to classifier the classifier will test the item sets with trained item sets. A distributed function is used to classify the minimal infrequent item sets and produce the minimal infrequent weighted item sets.



This module describes about the training of the item sets with the SVM classifier. The infrequent item set is trained and the infrequent item sets are entered into the classifier. The classification of minimal infrequent weighted item sets is done by the distributed function.

SVM MODEL IMPLEMENTATION

A supervised binary classification problem [8] is considered for finding minimal item sets. If the training diabetic data are represented by $\{xi, yi\}$, $i = 1, 2, \ldots, N$, and $yi \in \{-1, +1\}$ where N is the number of training samples i.e., minimal infrequent item sets, $yi=+1$ for class $\omega 1$ that represent the weighted frequent item sets and $yi=-1$ for class $\omega 2$ that represent the weighted infrequent item sets. the two classes are linearly

separable i.e., it is possible to find at least one hyper plane i.e., one of the assigned weighted vector value defined by a vector w i.e., input vector value with a bias w0 interval between the training class weight, which can separate the test class without error: The below equation to find the classification value

$$\boldsymbol{f(x) = w.\, x + w0 = 0}$$

To find such a hyper plane, w and w0 should be estimated in a way that $y_i(w\cdot x + w) \geq +1$ *i* for yi = + 1 (class ω1) and $y_i\ (w\cdot x + w) \leq -1$ for yi = - 1 (class ω2). These two, can be combined to provide more hyper plane: $y_i(w\cdot x + w) -1 \geq 0$. Many hyper planes could be fitted to separate the two classes but there is only one optimal hyper plane that is expected to generalize better than other hyper planes. The goal is to search for the hyper plane that leaves the maximum margin between classes i.e., maximum difference between frequent and infrequent item sets. To be able to find the optimal hyper plane, the support vectors must be defined. The support vectors (training dataset of infrequent item sets) lie on two hyper planes which are parallel to the optimal and are given by: $w.\, x + w0 = \pm 1$. Maximize the feature space using w, w0 and ratio of these values. Find optimal hyper plane using Lagrangian formulation. Press stated that [6] the optimal hyper plane discriminate function becomes:

$$f(x) = \sum \lambda_i y_i (x_t x) + w_0$$

$$(i \in S)$$

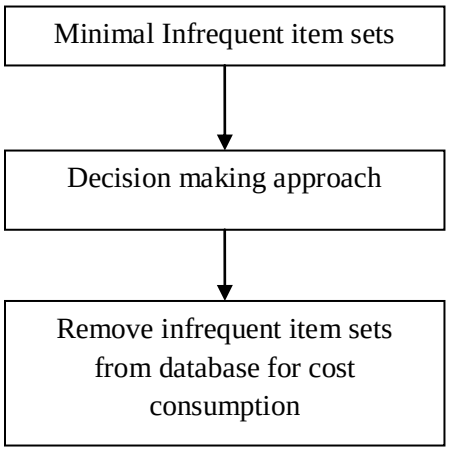
where λ_i is the Lagrangian multipliers, S is the subset samples i.e., infrequent weighted item sets. The linear classification is not possible to classify the testing data thus we need some cost function to estimate the error between training sample dataset and testing data. Thus some kernel function can be applied to estimate cost between the training data and testing data. The final result value can be computed by following equation:

$$f(x) = \sum \lambda_i y_i K (x_i x) + w_0$$

$$(i \in S)$$

where $K(x_i, x)$ is the kernel function.The kernel function may be a linear function or a polynomial function or a radial basis function or a tangent function or sigmoid function

DECISION MAKING SYSTEM



In this stage the minimal infrequent item sets are found and a decision making approach is made by the user to remove the item sets from the database. The purpose of removing the item sets is to minimize the cost consumption. Finally, the minimal infrequent weighted item sets are classified and the decision making approach algorithm is used to remove the minimal infrequent weighted item sets from the database in which it is stored.

5. RESULTS AND DISSCUSSION

In this section, a thorough discussion about the result and performance measures of the proposed system is discussed. This approach can be used in online purchasing of products. If the user wants to buy an item first then the weighted value should be noted. If the product has more weight value then it is sold frequently. From the weight value the user can determine whether to buy the particular product or not.

In database, if an item set is not used by users then the cost for storing that item set in database is useless. In order to reduce the cost function of user this minimal infrequent item sets are found-out and based on the weight value the user make decision whether to store or eliminate the item set in the database.

6. CONCLUSIONS AND FUTURE WORK

The decision making system can implement some features to the SVM machine. The SVM machine with applied features will automatically generate the minimal infrequent weighted Item sets. This method is easiest method to identify the minimal infrequent weighted item sets and the user process these sets to keep or not in order to reduce the cost consumption. The efficiency and time complexity is also very high compared to other results. The future work of this project is to convert the infrequent weighted item set to the frequent weighted item sets.

REFERENCES

[1] Luca Cagliero and Paolo Garza, "Infrequent Weighted Itemset Mining Using Frequent Pattern Growth,"IEEE Trans. Knowledge and Data Eng.,Apr.2014, vol. 26, no 4, pp. 903 – 915.

[2] http://en.wikipedia.org/wiki/Data_mining

[3] http://docs.opencv.org/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

[4] D.J. Haglin and A.M. Manning, "On Minimal Infrequent Item set Mining," Proc. Int'l Conf. Data Mining (DMIN '07), pp. 141-147, 2007.

[5] K.Sun and F.Bai, "Mining Weighted Association Rules Without Preassigned Weights," IEEE Trans. Knowledge and Data Eng. , Apr. 2008, vol. 20, no. 4, pp. 489-495.

[6] C.K.-S. Leung, C.L. Carmichael, and B. Hao, "Efficient Mining of Frequent Patterns from Uncertain Data," Proc. Seventh IEEE Int'l Conf. Data Mining Workshops ICDMW '07),2007, pp. 489-494.

[7] Press, William H.; Teukolsky, Saul A.; Vetterling, William T.; Flannery, B. P. (2007). "Section 16.5. Support Vector Machines". Numerical Recipes: The Art of Scientific Computing (3rd ed.). New York: Cambridge University Press.

[8] Xiaojin Zhu , Zoubin Ghahramani , John Lafferty,School of Computer Science, Carnegie Mellon University, Pittsburgh PA 15213, USA ,Gatsby Computational Neuroscience Unit, University College London, London WC1N 3AR, UK "http://www.aaai.org/Papers/ICML/2003/ICML03-118.pdf"