ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(22), DOI: 10.17485/ijst/2015/v8i22/79185, September 2015* ISSN (Online) : 0974-5645

A Weighted Distance Metric Clustering Method to Cluster Small Data Points from a Projected Database Generated from a FreeSpan Algorithm

S. Gayathri[1*], M. Mary Metilda[2] and S. Sanjai Babu[1]

[1]Bharathiar University, Coimbatore - 641046, Tamil Nadu, India; gay_sri123@yahoo.com, ssanjai2000@gmail.com
[2]Queen Marys College, Chennai - 600004, Tamil Nadu, India; metilda_dgvc@yahoo.co.in

Abstract

Background/Objectives: Clustering and Sequential Pattern Mining is two most important unsupervised learning algorithms. The objective is to mine small projected databases rejected by Frequent Pattern - Projected Sequential Pattern mining (FreeSpan) technique using a weighted distance metric clustering method, a process of finding the distance between the small data points and cluster it so that it cannot be rejected. **Methods/Statistical Analysis:** The method involves the implementation of a distance metric clustering algorithm over a FreeSpan technique to cluster the data points of small projected databases. The FreeSpan technique can be considered as an ensemble of clustering and sequential pattern mining methods. **Findings**: The clustering method clusters the data points resulted from the FreeSpan technique that are ignored after the scanning process as their sizes are very small. The clustered data therefore gathers the ignored data points thereby providing an accurate clustered data containing small data points which results is trustable sequential pattern for future predictions. The proposed system reduces the complexity by incorporating just a single clustering algorithm. Therefore the major operations of the algorithm remain undisturbed and give its efficient output and also the output is found to be accurate and stable. **Applications/Improvements:** The technique proposed in the paper can be applied to datasets that needs to be clustered for decision making. The same technique holds good and can be made applicable to high dimensional views.

Keywords: Clustering, Distance Metric Method, FreeSpan, Projected Databases

1. Introduction

The Research paper explores data mining concepts. The data mining is the process of discovering useful knowledge from the data where data mining is the specific important step in the process[1]. The paper uses the two most important techniques of data mining which are the earliest techniques of data mining. Sequential pattern mining is one among the two algorithms. The operation of the sequential pattern mining is to discover all the sequential patterns with a minimum support specified by the user where the number of data sequences that contain the pattern forms the support of a pattern[2]. It is proved that the sequential patterns are efficient in handling large databases to do an incremental mining[3]. Many varied forms of sequential pattern mining have been proposed. One such technique is Frequent Pattern-Projected Sequence pattern mining. The reason behind using this kind of sequential pattern is in normal sequential pattern mining the projected database of small points are rejected as outliers but actually these rejected data points so to avoid this set back, the FreeSpan algorithm is used which is an efficient technique in creating the sequential pattern from the projected databases of small points. Though the FreeSpan is an efficient algorithm it faces issues when handling multilevel information that results in issues

*Author for correspondence

with future predictions of sequence. So to make a reliable future prediction a clustering method called weighted distance metric method is used. The algorithm scans each and every object in the projected database and finds the distance between the object and a seed[4]. Then with the data point obtained are clustered and separated. So when a prediction is made it can be found that cluster with such data points predicts their respective patterns. The detailed view about the algorithm is given below.

The General Sequential Pattern (GSP), mining method is based on the Apriori algorithm has few cons like slow processing and delayed response. When the method is used in the projection of databases the GSP algorithm would not mine smaller projected databases. Generally every sequential pattern mining technique follows a basic ground rule that any super pattern of an infrequent pattern cannot be frequent[1]. Based on the heuristic, many sequential pattern mining approach have also been achieved. Considering the databases similar to the case mentioned above, the sequential pattern mining is even used for large sequence database. The large sequence databases have the ability to generate huge set of candidate sequences. The reasons for such huge outcome of sequences are due to iterations of items in a sequence in all possible permutations which results in huge data outcomes that cannot be ignored. The main drawback in the usage of GSP algorithm at larger sequential pattern the algorithm becomes inefficient at a point because Apriori algorithm is a step-by-step process which makes it very difficult to mine such huge set of data within the expected time. Therefore to achieve a high active performance to search throughout the projected databases J. Han developed a mining approach which was faster than GSP and also was able to mine smaller projected databases to even a longer processing time period.

The FreeSpan mining method in short can be defined as the usage of frequent items to recursively project sequential databases into smaller projected databases and generate subsequent fragments in every projected database. The reason behind the technique being used in the paper is the algorithm, far more reliable and efficient than the GSP algorithm though the technique pattern clustering has been experimented in many papers it is not yet used in FreeSpan approach. Since the scanning process contains different outcomes for various iterations the resultant projected database has to face the consequence of complexity in estimating future predictions using these longer sequential patterns thereby resulting in failed predictions. In order to overcome such constraints we use a clustering technique which involves weighted distance measurement.

The paper is organized as follows. Section II contains the methodology used for the research work namely FreeSpan algorithm. The implementation of the algorithm is discussed in section III. Section IV explores about the results and discussion and finally, V concludes the research work.

2. Methodology

There are many related works that conceptualize the idea of incorporating a clustering algorithm with a sequential pattern. Shuai Ma proposed the combination clustering with moving sequential patterns[5]. Dilan Perera developed the technique of online collaborative of learning data by combining clustering and sequential pattern mining[6]. Eric Hsueh-Chan Lu proposed the technique of cluster based sequential patterns in spatial data mining for location based services[7]. The primary aim of any mining process would be the accuracy. One such way to get an accurate result is removal of noise by repeated processing of the data. The more the iteration takes place the more accurate is the data. At the same time in sequential pattern mining when the iteration is repeated the size of the sequence pattern also increases. To be noted, the time for deriving the projected databases from the set of frequent items is already known[2]. The repetition forms the basic rule of pattern growth approach. Sometimes, the same approach is applied to the projection of a database. Therefore when such an algorithm is applied to sequential pattern mining using the resultant items of a frequent pattern set, it is called Projected Sequential pattern mining[5].

The FreeSpan algorithm can be well explained using a set $I = \{i_1, i_2 \ldots i_n\}$ where the elements inside I are called items and the subset of I are called item set. In general a sequence can be defined as an orderly arrangement of elements and in the case a sequence can be represented as $S = <s_1, s_{2\ldots\ldots} s_l>$[4]. For the set I, s_j is an item set which means $s_j \subseteq I$ and for reference s_l is also called as the element of the sequence which contains $(x_1, x_2 \ldots\ldots, x_m)$ and as far as this set is concerned x_k is an item which means $x_k \epsilon$ I. As a general rule an item can occur at least once in the element of a sequence. Now let us consider two sequence α and β where $\alpha = < a_1, a_2, \ldots . a_n >$, $\beta = <b_1, b_2, \ldots .. b_m>$, here α is the subsequence of B so this means β is the super sequence,

which implies $\alpha \subseteq \beta$ where $1 \leq j_1 < j_2 < \ldots j_n \leq m$ such that $a_1 \subseteq b_{j1}$, $a_2 \subseteq b_{j2}, \ldots, a_n \subseteq b_{jn}$. The next process is the update of the sequence into the sequence database S. The sequence database is generally represented as a pair of values as $<s_{id}, s>$. If the sequence has $\alpha \subseteq s$ then the sequence database contains a sequence α. The tuples in the database is determined by defining a support threshold value ϵ. If the α, is ϵ number of tuples in S, then the sequence α is a sequential pattern.

The sequential pattern can be found by the following ways. Initially the sequence database S with minimum support 2 is processed. The process begins with scanning the sequence database once. The set of frequent items obtained by scanning is denoted as L1. The resulting set is a set of length-1 sequential patterns. These patterns are arranged in a descending order of support. The frequent items are represented as < b:5, c:4, a:3, d:3,e:3, f:3 >. At present there are 6 items formed after first scanning process. These lists of frequent items are also called as f_ List. The sequential patterns to be mined make use of the data provided in the sequential database columns shown in Table 1.

The following step is to divide the sequential patterns to six subsets without overlapping any of the elements in the list. The dividing process is done as follows:

- The sequence of those having the item f.
- Those having the item e but not f.
- Those having the item d but neither e nor f.

The process is repeated for further sequences and the method is called Divide-and-Conquer. Once the sequential patterns, are divided a Frequent Item matrix is created. The purpose of creating the matrix is to count the number of occurrence of frequent item of length-2 sequence formed by the f_list. The Frequent Item Matrix of the f_list$< i_1, i_{2..} i_n>$ is a triangular matrix of F[j,k] where $1 \leq j \leq m$ and $1 \leq k < j$ for every $1<j<m$. Only one counter takes place for every single scanning of F[j,j] but for F[j,k]three counters namely U,V,W takes place. The three counters can be explained in detail as,

U → number of occurrences of ik after ij, which can be represented as an <ij, ik> sequence.

V → number of occurrences of ik before ij, which can be represented as an <ik ,ij> sequence.

W → number of occurrences of ik which can be represented as an <(ij, ik)> sequence.

In the above illustration the f_list has 6 items <b:5, c:4, a:3, d:3,e:3, f:3> therefore 6 x 6 triangular frequent matrix F with every single counter initialized from 0. The matrix is filled up during the second scan. The first generated sequence <(bd)cb(ac)> increases the first two counters of matrix by a value of 1 which can be shown as F[b,c] = (1,1,0), this is because only two <bc> and <cb> are considered but not <bc>. Nearing to F[b,d] as only <(bd)> and <db> occurs in this sequence the value of F[b,d] = (0,1,1). The process is continued until there are no objects or items available to scan.

- As mentioned earlier the frequent item matrix is used to generate the length-2 sequence pattern.
- The set of projected database is used to generate length-3 sequence patterns or even longer patterns.

Let us consider an item set X. For the item set, the projected database should have all the items in X. All the infrequent items and the items after X are discarded. Similar to the case of X item set the α-projected database is a collection of sequences that have the items of α as a subsequence. The next task is to generate the third level database. The process involves creating a set of Annotations to indicate the set of items or a sequence that must be examined in the projection and later for mining level-3 databases. Let us check the usage of matrix to generate the following:

- length-2 sequential patterns.
- Annotations of item repeating patterns.
- Annotations of projected databases.

The generation of length-2 sequence pattern is explained as for each counter if the value of the counter is not less than the minimum support then it produces the output of the corresponding frequent pattern that is already seen in the current paper.

Table 1. Scanned sequence and frequent item pattern

Sequence_id	Sequence	Frequent item pattern
10	<(bd)cb(ac)>	{a,b,c,d}
20	<(bf)(ce)b(fg)>	{b,c,e,f,g}
30	<(ah)(bf)abf>	{a.b,f,g}
40	<(be)(ce)d>	{b,c,d,e}
50	<a(bd)bcb(ade)>	{a,b,c,d,e}

The annotations of item repeating pattern uses either <> - a particular order of sequence or {} - a sequence of any order. It also looks for more than one occurrence of a_i which is denoted as a^+. The important thing to be noted is at least one of the a_i or a_j must repeatedly occur. The annotation of the projected database is a step by step counter process that takes place for every row and every column. The projected columns is determined after examining all columns in front of I then the annotation is given as output containing i, j and the set of projected columns. The process is explained in detail by J. Han et al.[4] in their work mentioned in the reference[1]. After generating the annotations the matrix can be discarded. The remaining mining process will be confined to smaller projected database. The set of item-repeating patterns, the projected database and their sequential patterns is given in the Table 2.

The FreeSpan algorithm gives the sequential patterns inside a projected database. The model is well utilized to predict the occurrences of some event (included in the data set) in the future and understand the intrinsic characteristics included in it[8], is the lead point for extending and enhancing the results of the freespan technique. Though the patterns obtained are satisfactory, it is not completely reliable at few cases like using longer sequential patterns for futuristic predictions. To retain a classified data and to avoid future failures in the sequence predictions and also to take user control over the data, the data is clustered as sequence patterns.

3. Algorithm-distance Metric Clustering

The conversion of multilevel information into a database often traps into an issue of constraint based sequential pattern mining. So a weighted distance metric clustering technique is used to form clusters of data[9]. Di Wu and Jiadong Ren proposed a method almost similar to this research work, but we use weighted distance metric clustering but they have used weighted sequential pattern clustering for sequential pattern mining[10]. In sequence clustering a portion of the sequence S = {s_1, s_2, s_n} is mentioned as a 'segment'. The objective of the paper is to categorize the sequence into clusters in accordance with their sequential similarities. A typical pattern clustering involves definition of a pattern proximity measure appropriate to the data domain[11], which forms the basis of our approach. The research paper uses an algorithm for partitioning the object into clusters by applying a threshold. To begin with, consider two objects related to each other at a specific distance. Since the distance is the origination spot of the process the value is called as a seed point. Let us consider two clusters as two objects which are taken into account. The scanning process is then initiated to estimate the distance between every objects in the cluster. Prior initiating the process a threshold value δ is calculated by finding the average distance between the objects. During the scanning process the objects at a distance of less than or equal to the threshold value δ are clubbed into a cluster. The objects that have distance greater than δ do not belong to the cluster. Now, there is a need to find the second seed. A second scanning is made at the objects outside the cluster or technically called as outliers to find the minimum distance objects and the second seed is formed. The process is iterated till there exists with no object in the sequence to be clustered.

3.1 Algorithm

Step 1: Set of items in a se I= {i1, i2,…in}.

Step 2: Set S = {s1,s2,…sj}.

Step 3: Let β = <b1,b2,…bn>, β is the super sequence of α.

Table 2. The projected database and its sequential patterns

Annotation	**<(ce)>: {b}**	**<da>:{b,c}**	**{cd} : {b}**	**<ca>:{b}**
Projected Database	<b(ce)b>, <b(ce)>	<b(ce)b>, <b(ce)>, <(bd)bcba>	<(bd)cbc>, <bcd>,<(bd)bcbd>	<bcba> <bbcba>
Sequential Patterns	<b(ce)> : 2	<(bd)a>:2, <dca>:2, <dba>:2, <(bd)ca>:2, <(bd)ba>:2, <dcba>:2, <(bd)cba>:2	<bcd>:2, <(bd)c>:2, <dcb>:2, <(bd)cb>:2	<bca>:2, <cba>:2, <bcba>:2

Step 3: Let α = <a1,a2,…an> , α is the subsequence of β.

Therefore, α ⊆β and a1⊆ bj1, a2 ⊆bj2……. an⊆ bjn.

Step 4: If α is ϵ number of tuples in S then the sequence α is a sequential pattern.

Step 5: Initiate – Scanning process for min_threshold 2.

Step 6: Scanned sequence contains 6 items → F_List.

Step 7: Divide and Conquer rule. Generate a frequent item matrix, compute the number of

Occurrences for length-1 sequence.1

Step 8: Repeat – Generate a frequent matrix, compute the number of occurrences for length-2

Sequence and repeat for length-3 sequence.

Step 9: Annotations of projected databases- Find the set of item repeating patterns form the data

of step 7 and 8.

Step 10: Distance metric algorithm. Input: The threshold distance δ.

Step 11: Compute the distance between the first two objects in the sequence pattern.

Step12: Cluster the objects whose distance ≤ δ. Remove the objects as outliers whose

Distance > δ.

Although Haixun Wang et al have proposed a clustering model based called pCluster to capture both the closeness and similarity of patterns shown by the objects[12]. The technique does not serve the purpose for smaller projecting sequence in projected databases. Also Eric P. Xing et al used a distance metric model for clustering data[13]. We considered (a, b, c, d, e, f and g) as a training dataset. Out of these four items a, b, c and d are finally sequenced by their frequency of occurrence at every counter. The sequential patterns of the four projected databases are listed in Table 3. The four sequences can be represented as A, B, C and D. When the minimum value of the respective scanned distance in the tabular column is less compared to the highest calculated distance value then these sequences will merge into a cluster. In all consecutive iteration new clusters are formed according to their distance measurement. These clusters formed can be used for future predictions without any constraints in the sequential patterns.

4. Results and Discussion

The results of the proposed paper are determined by the outcome of two algorithms which has the results of both the FreeSpan technique and weighted distance metric technique. The research begins from a set I with elements $\{i_1, i_2 \ldots i_n\}$ called item and the order they have been arranged is called a sequence. In a database the sequence is represented as a pair and the database is called sequence database. When saying database it has tuples in it. The tulpes are set by a support threshold ε then it is said to be a sequential pattern$<s_{id}, s>$. If the values in the sequence are more than ε then it is a sequential pattern. Every set has a subset in FreeSpan technique. The sequential pattern finding method begins by giving different minimum support like 2, 3 ad so on. The frequent pattern mining is used in sequence pattern mining so that a frequent item matrix can be formed from a frequent item set. The frequent pattern mining is well explained in[14] by Charu C. Agarwal and J. Hann and the approach for a frequent pattern mining is well explained in[15] by Jiawei Han et al.[4] The scanning of the sequence database gives the first set of sequence patterns called the length-1 pattern which are arranged in descending order which is denoted in Table 3.

The advantage of this Freespan technique when compared to other is it scans the first set of sequence and holds on the output and starts the process again. Non-overlapping data are very important in case of

Table 3. The distance between the points in a sequence database

Sl.No	A	{B,F}	C	D	E	G
A	0	3.2	4	4.5	5	5.5
{B,F}	4.5	0	5.5	9.5	8	3
C	3	5.5	0	5	10	3
D	5	9.5	5	0	12	5
E	4	8	10	12	0	4
G	6	3	3	5	4	0

data mining as they give unique results. So the items in a result are confirmed in such a way that an element once appeared in a sequence is not repeated with the same combination. A frequent item matrix is also created in which the number of occurrences takes place in the way a sequence pattern repeats, giving the source for the predictions or future behavior of a sequence. The technique serves good for a single level information prediction but for multi level information into a database faces accuracy issues. So to overcome this, a weighted distance metric clustering technique is used to cluster the items in the projected sequence into clusters. This is done because when the sequences are clustered, the role of each sequence items can be known so that it would be easy to predict that a sequence containing such an item would result in that specific sequence pattern. Table 3 shows the metric distance between all the different data points in a projected database. The δ threshold value is the input distance parameter and the points that are at a distance lesser than δ are clustered together. The points that are separated at a distance more than the δ value are considered as outliers.

Clustering can be done in many ways but this weighted distance metric is found to be an apt one because the complexity of the algorithm is reduced when clustering is used compared to other technique. Any clustering item should contain a list of data and its values, so here the sequence items are taken as data. There values are partitioned by a distance metric approach with the help of a threshold value δ. If the distance is lesser than the threshold value the objects in the sequence are clustered and if the distance is more it is considered as an outlier. The clustering is repeated for every scanning process and the output is retained. The performance shows that the mining of smaller projected database becomes more efficient when a clustering method is involved in the scanning process. The clustering method is tested using the weka tool kit and the datasets are taken from the real world data sets from the UCI Machine Learning Repository. The test is done with an Intel processor environment with a 2 GB RAM. The sequential pattern mining along with the weighted measure technique for clustering are tested using weka tool kit.

5. Conclusion

The research paper involves the combination of two efficient algorithms to find a solution for reliable sequential pattern predictions. Freespan technique is considered as a pioneer in sequential pattern mining which combines the idea of frequent item and sequence patterning in a projected database. The significant advantage of the technique is that it processes even the smaller projected databases without neglecting the small sequences unlike other techniques. Clustering these refined patterns involves novel approach. Though the technique is proposed earlier it has not been used with freespan algorithm. The process has been found to be more efficient and less complex. The approach can further be extended in future by involving time constraints and proposing the same approach for high dimensional dataset.

6. References

1. Alijamaat A, Khalilian M, Mustapha N. A novel approach for high dimensional data clustering. Proceedings of third International Conference on Knowledge Discovery and Data Mining; 2010. p. 264–7.
2. Olson DL, Delen D. Advanced data mining techniques. Berlin-Heide: Springer Science and Business Media; 2008.
3. Cheng H, Yan X, Han J. IncSpan: Incremental mining of sequential patterns in large database. Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2004. p. 527–32.
4. Han J, Pei J, Yan X. Sequential pattern mining by pattern-growth: Principles and extensions. Foundations and Advances in Data Mining. Springer Berlin Heidelberg; 2005. p. 183–220.
5. Ma S, Tang S, Yang D, Wang T, Han J. Combining clustering with moving sequential pattern mining: A novel and efficient technique. Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg; 2004. p. 419–23.
6. Perera D, Kay J, Koprinska I, Yacef K, Zaiane OR. Clustering and sequential pattern mining of online collaborative learning data. IEEE Transactions on Knowledge and Data Engineering; 2009 June 21. p. 759–72.
7. Lu EHC, Tseng VS. Mining cluster-based mobile sequential patterns in location-based service environments. Tenth International Conference on Mobile Data Management: Systems, Services and Middleware; 2009. p. 273–8.

8. Pei J, Han J, Mortazavi-Asl B, Pinto H, Chen Q, Dayal U, Hsu M-C. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. Proceedings of IEEE International Conference on Data Engineering; 2001. p. 215.
9. Chezhian V, Thanappan SU, Ragavan SM. Hierarchical sequence clustering algorithm for data mining. Proceedings of the World Congress on Engineering. 2011; 3:223–34.
10. Wu D, Ren J. Sequence clustering algorithm based on weighed sequential pattern similarity. Telkomnika Indonesian Journal of Electrical Engineering. 12(7):5529–36.
11. Pei J, Mortazavi-Asl B, Wang J, Pinto H, Chen Q, Dayal U, Hsu MC. Mining sequential patterns by pattern-growth: The prefixspan approach. IEEE Transactions on Knowledge and Data Engineering. 2004; 16(11):1424–40.
12. Haixun W, Wang W, Yang J, Yu PS. Clustering by pattern similarity in large data sets. Proceedings of the 2002 ACM SIGMOD international conference on Management of data; 2002. p. 394–405.
13. Xing EP, Jordan MI, Russell S, Ng AY. Distance metric learning with application to clustering with side-information. Advances in Neural Information Processing Systems; 2002.
14. Aggarwal, CC, Li Y, Wang J, Wang J. Frequent pattern mining with uncertain data. Proceedings of 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2009. p. 29–38.
15. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. Proceedings of the 2000 ACM SIGMOD Record. 2000; 29(2):1–12.