RESEARCH ARTICLE OPEN ACCESS

Probabilistic Deterministic Classifier Based Sequential Pattern Mining to Evaluate Structural Pattern on Chemical Bonding

S.Sathya [1], N.Rajendran [2]
Research Scholar [1], Bharathiar University, Coimbatore
Principal [2],Vivekanandha arts & science college,Sankiri,Salem(dt)
India

ABSTRACT

Evaluating the structural patterns of chemical bonding involves identifying subset of the most probable chemical bonds that produces compatible results through analysis of sequential patterns. A sequential pattern mining algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to classify structure of chemical bonds, the effectiveness is related to the number of matched sequential patterns. Based on these criteria, a Probabilistic Deterministic Classifier and Sequential Pattern Mining (PDC-SPM) is proposed and experimentally evaluated in this paper. The PDC-SPM evaluates the structural patterns of chemical bonding in chemical information data sets. Naive Bayes Probabilistic Deterministic Classifier identifies the structures of chemical bond with training samples. Probabilistic Deterministic Classifier provides probability of membership in each class of chemical bonds for identifying the new bonds. With the identified structures of new bonds, Sequential Pattern Mining is applied to evaluate structural patterns of chemical bonding. Based on these criteria a Sequential Pattern Chemical Structure Bond Mining (SPCSBM) algorithm is proposed and experimentally evaluated in this paper. The efficiency and effectiveness of the SPCSBM algorithm are evaluated through an empirical study. Extensive experiments are carried out and the performance of PDC-SPM is evaluated and compared against existing state-of-art techniques. The results show that the technique not only improves matched sequential pattern rate but also reduces the chemical bond classification time.

***Keywords*:-** Chemical bonding, Sequential pattern mining, Structural patterns Probabilistic Deterministic Classifier, Membership

I. INTRODUCTION

Chemical bonding is the interactions that account for the association of atoms into molecules, ions, crystals, and other stable species that make up the familiar substances in chemical information. Recently novel method enriching chemical bond classification and patterns mined have been evolved.

Probabilistic Sequence Translation Alignment Model (PSTAM) [1] for time series classification using probabilistic model in order to efficiently capture the class specific patterns. Despite improvements observed in classification accuracy, the time at which the classification was performed remained unaddressed. A cluster based feature selection algorithm (FAST) [2] concerning time required to classify was presented using Minimum Spanning clustering tree method. A survey on frequent sub-graph mining algorithms was presented in [3].

Many researchers have published their study of eventually mining the patterns from different applications. In [4], a methodology for interactive mining using interactive visualization technique was presented to support ad hoc visual exploration of patterns mined for health record data. Probabilistic models were applied in [5] to study the hydrogen bond stability resulting in early prediction of H-bond stability with the aid of tree construction.

For several decades, the major sources for clinical drug discovery lie in the study of Natural products (NPs). Proper mining through natural products results in identification of potential compounds. In [6], multiple compound mining approaches were applied to identify potential pyrromalide compounds during fermentation. In [7], RNA secondary structure prediction with the aid of soft computing was presented using metaheuristics.

With the problem of sequential pattern mining proposed by different authors due to its wide variety of applications, several methods were presented to improve the patterns to be mined. A new sequential pattern mining algorithm called, Node Linkage Depth First Traversal

algorithm [8] was presented to improve the rate of scalability. In [9], patterns in bio chemical reaction were mined using a tool chain that identified reoccurring patterns improving the rate of retrieval. However, the selection bias issue that patterns were selected from large number of candidates in database remained unaddressed. To solve this issue, in [10], a framework called selective inference was presented for finding scientifically significant patterns from database.

Based on the aforementioned techniques and methods, in this paper, a Probabilistic Deterministic Classifier and Sequential Pattern Mining technique is presented. The paper is organized as follows: In section 2 some related classification and pattern mining methods are briefly reviewed. After formally stating the problem, our main approach is described in Section 3 with the detailed learning and inference algorithms. In Section 4 experimental settings are presented with which an extensive set of evaluations is conduction in Section 5. Section 6 provides conclusion.

II. RELATED WORKS

In most existing pattern mining methods, the consistent aspect of the interaction network such as neighbour counting, network analysis, graph pattern mining were concerned. However, application of gap constraints given with the sequential pattern mining remained unaddressed. In [11], to overcome the inconsistency problem, a tree graph was constructed with the aim of obtaining most accurate prediction with the highest accuracy rate. In [12], a concise border like presentation was presented using itemset based distinguishing sequential patterns to improve the efficiency.

With the discovered knowledge using frequent sub-sequences, sequential-pattern mining was used to discover the subsequences from sequence database. Many research scholars have only handled the static database to identify the desired sequential patterns. In [13], prelarge concept was adopted for efficient handling of discovered sequential patterns with sequence deletion aiming at reducing the execution time. A new domain of patterns called asubtrees was introduced in [14] with the aid of itemset closure and structural closure resulting in the good compromise between non-redundancy of solutions and execution time.

Sequential pattern mining being computationally challenging, several researchers have applied several methods to reduce the complexity involved in the task of mining closed sequential patterns. In [15], a novel algorithm for mining closed frequent sequences of itemsets combining new data representation of the dataset, based on sparse id-lists and vertical id-lists was presented to reduce the pruning time and improve the search space. However, temporal relationships remained unaddressed. To solve this issue, in [16], to mine sequential patterns of diabetes a method that identified temporal patterns. Another method based on temporal patterns was presented in [17] using first order Markov behaviour to demonstrate the feasibility and effectiveness of the approach.

One of the new frontiers in the field of communication engineering and networking is molecular nanonetworks based on the paradigm called Molecular Communication. In [18], a probability distribution model using Power Spectral Density (PSD) was applied to reduce the inference rate. In [19], hybridization of data mining and natural language processing was designed to improve the gene interaction detection rate.

III. METHODOLOGY

Chemical bonding is the interactions that account for the association of atoms into molecules, ions, crystals, and other stable species that make up the familiar substances in chemical information. Let us represent the compounds for chemical bonding be in a form of graph '$G = (V, E)$', where 'V' corresponds to the vertices with atom element type whereas the edges 'E' with the bond type are in undirected as there exists no directional associative with chemical bonds. Figure 1 shows the framework of the proposed Probabilistic Deterministic Classifier and Sequential Pattern Mining (PDC-SPM).

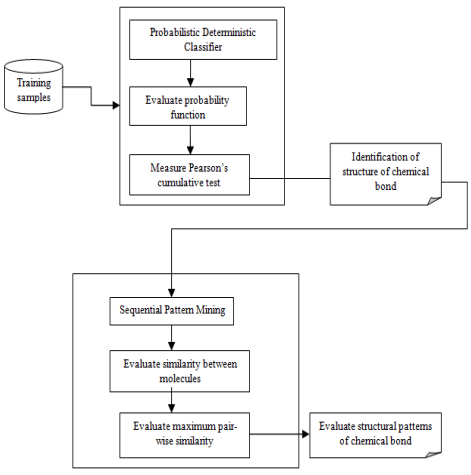

Figure 1 Framework of the proposed Probabilistic Deterministic Classifier and Sequential Pattern Mining

Chemical compounds are well-defined structures that are easily compressed by a graph representation. Improper classification of chemical compounds, along with irrelevant patterns, severely affects the number of matched sequential patterns [1] [2]. Thus, a classifier method selection should be able to identify the structure of chemical bonds in a significant manner and evaluate more relevant structural patterns as possible.

Keeping these in mind, we develop a novel technique which can efficiently and effectively deal with both proper classification and evaluate relevant structural patterns and obtain a good sequence patterns for identified chemical bonds. We achieve this through a new Probabilistic Deterministic Classifier and Sequential Pattern Mining technique as shown in figure 1 with two connected components that used Probabilistic Deterministic Classifier and Sequential Pattern Mining model. The elaborate description is presented in the following section.

A. Probabilistic Deterministic Classifier

Initially, structures of chemical bond are identified with training samples using Probabilistic Deterministic Classifier with the aid of Naive Bayes model. Let us assume that that the chemical properties of each atom in a bond be defined by a probability. Therefore the whole bond in the proposed technique is defined by set of two probabilities and is as expressed below.

$$\sum_{i=1}^{2} Prob_i \qquad (1)$$

Let us further consider that the total valence electrons are disseminated between two atoms, then, the probability function is expressed as given below.

$$Prob_i = \frac{(V_O)_i}{V_E} \qquad (2)$$

From (2), 'V_O' corresponds to the number of valence occupancies whereas 'V_E' corresponds to the valence electrons and is as expressed below.

$$V_E = (V_O)_i + (V_O)_j, where\ V_E\ lies\ between\ [0,1]\ for\ Prob_i \quad (3)$$

The equation given in (2) defines normalized probability set. The proposed technique uses Naive Bayes Probabilistic Deterministic Classifier to identify the solution for measuring the closeness of these two probabilities '$(V_O)_i$' and '$(V_O)_j$' to each other. In the case of normalized probability set, pure covalent bond occurs whereas the ionic nature of the bond becomes more due to the increase in between the two probabilities '$(V_O)_i$' and '$(V_O)_j$' respectively.

The Naive Bayes Probabilistic Deterministic Classifier in the proposed technique introduced a hypothesis value of '0.5' for testing. With the introduced hypothesis, the Probabilistic Deterministic Classifiers provides probability of membership in each class of chemical bonds for identifying the new bonds that are more proven to be more useful than performed with classification rules using Person's cumulative test..

$$\chi^2 = \sum_{j=1}^{n} \frac{(O_j - E_j)^2}{E_j} \qquad (4)$$

From (4), the Pearson's cumulative test 'χ^2' for identification of structures of chemical bond is obtained by the ratio of square of difference between the observations 'O_j' of type 'j' and expected frequency 'E_j' of type 'j' to the expected frequency 'E_j' respectively. The Naive Bayes Probabilistic Deterministic Classifier uses training samples to identify the structure of chemical bonds. Figure 2 shows the example structure of chemical bonding.

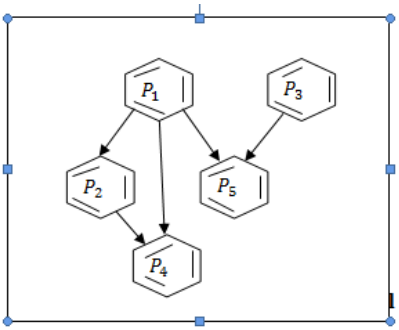

Figure 2 Example structure of chemical bonding (a sample graph)

From the figure, the dependencies between the nodes (i.e. atoms or ions) are measured from which the learnt structure of chemical bonds is then mined using sequential pattern mining that helps in the identification of highly discriminative patterns for in chemical information datasets.

The main assumption underlying the Naive Bayes Probabilistic Deterministic Classifier is that every attribute (in our case, atoms, molecules, ion, crystals) 'Att_i', is conditionally independent from the rest of the attributes, given the state of the class variable (in our case, the chemical bonds), 'Cl'. Therefore, Naive Bayes Probabilistic Deterministic Classifier for identification of structure of chemical bond with a finite set '$U = \{Att_1, Att_2, \dots, Att_n, Cl\}$' is formulated as given below.

$$Prob\,(Cl, Att_1, Att_2, \dots, Att_n) = Prob\,(Cl\,|Att_1, Att_2, \dots, Att_n)$$

$$= \alpha * Prob\,(Att_i \mid Cl) \qquad (5)$$

From (5), 'Att_i' corresponds to the attributes whereas 'Cl' represents the class variables with a constant factor 'α'. Thus the Naive Bayes Probabilistic Deterministic Classifier with Pearson's cumulative test obtains the closeness of two atomic probabilities '$(V_o)_i$' and '$(V_o)_j$' to the assigned hypothesis value of '0.5'. Therefore, the Naive Bayes Probabilistic Deterministic Classifier provides efficient means for probability of membership in each class of chemical bonds. Figure 3 shows the Naive Bayes Probabilistic Deterministic Classifier algorithm.

Input: Training samples Dataset 'DS',
Output: Efficient identification of structure chemical bonds
1: Begin 2: For each training samples 'DS' 3: Measure the probability function using (2) 4: Identify structure of chemical bonds using (4) 5: Evaluate Probabilistic Deterministic Classifier using (5) 6: End for 7: End

Figure 3 Naive Bayes Probabilistic Deterministic Classifier algorithm

As shown in the figure, the Naive Bayes Probabilistic Deterministic Classifier algorithm identifies the structure of chemical bond whose vertices and edges have been labeled with several descriptors, such as atom and bond types. To identify the structure of chemical bond, the probability function is measured with which the Pearson's cumulative test is performed to, such that the overall probability is maximized with respect to all possible alignments. Structure of chemical bonds is then measured using probabilistic function to perform efficient identification of structure chemical bonds.

B. Sequential Pattern Mining

From the identified structured chemical bond, Sequential Pattern Mining (SPMP is presented that evaluates the structural patterns of chemical bonding in chemical information data sets. These structured chemical bonds are then used to augment a maximum chemical bond function that calculates the maximum pair-wise chemical bond similarity between molecules. With this, different sequence patterns are identified and mined for the identified chemical bonds. Figure 4 shows the block diagram of Sequential Pattern Mining-based Chemical Structure Bond.

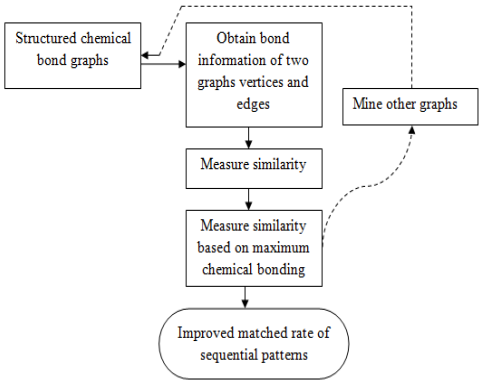

Figure 4 Block diagram of Sequential Pattern Mining-based Chemical Structure Bond

As shown in the figure, the Block diagram of Sequential Pattern Mining-based Chemical Structure Bond identifies the frequently occurring structured chemical bond groups represented as graphs, and measures the graph similarity measure based on the chemical bonding. We then build a Sequential Pattern Mining model to evaluate the structural patterns of chemical bonding.

With the significant application of SPM in the proposed technique, different sequence patterns provided for identified chemical bonds are obtained. Let us consider two graphs 'G_i and G_j' with vertices 'V_i ,V_j' and edges 'E_i, E_j' respectively. The similarity bond equation that computes the bond information of two vertices (atoms) is represented as given below.

$$SIM_{bond}\,(V_i, V_j) = \frac{1}{(V_i)(V_j)} \sum_{i,j=1}^{n} [bond\,(V_i), bond\,(V_j)] \quad (6)$$

The ' P_{bond} ' function, calculates the similarity between two vertices using their maximum chemical bond, is as given below.

$$P_{bond}(V_i, V_j) = P_v(V_i, V_j) + P_e(V_i, V_j) + SIM_{bond}(V_i, V_j) \quad (7)$$

The functions 'P_v' and 'P_e' measures the similarity between the atoms present in vertices and bonds in edges respectively. The resultant ' P_{res} ' of similarity of two chemical bonding graphs 'G_i' and 'G_j' is obtained as given below where the maximum chemical bonds between the vertices 'V_i' and 'V_j' are used

$$P_{res}(G_i, G_j) = MAX\ P_{bond}(V_i, V_j) \quad (8)$$

From (6), 'V_i' denotes the vertices in graph 'G_i' and 'V_j' denotes the vertices in smaller chemical bonding graph 'G_j' respectively. Figure 5 shows the Sequential Pattern Chemical Structure Bond Mining algorithm.

Input: Training samples Dataset 'DS', Vertices (represents the atoms) 'V_i', 'V_j', edges 'E_i, E_j' bonding graphs 'G_i' and 'G_j',
Output: efficient structural pattern mining
1: Begin 2: For two vertices 'V_i' and 'V_j' 3: Measure bond information of two vertices using (6) 4: Measure similarity of two chemical bonding graphs using (7) 5: Measure similarity between two vertices using their maximum chemical bond (8) 6: End for 7: End

Figure 5 Sequential Pattern Chemical Structure Bond Mining algorithm

As shown in the figure, the Sequential Pattern Chemical Structure Bond Mining (SPCSBM) algorithm consists of three steps (i) measuring bond through vertices (ii) constructing chemical bonding graphs and (iii) measuring similarity between vertices using maximum chemical bond. For a training sample Dataset ' DS ', with '$V = V_1, V_2, \ldots, V_m$' atoms and '$E = E_1, E_2, \ldots, E_n$' bonds, the first step measures the bond information for two vertices. In the second step, similarity of two chemical bonding graphs between the vertices and edges is measured. After constructing the chemical bonding graphs, in the third step, similarity between vertices are measured with the aid of maximum chemical bond. This in turn improves the number of matched sequential patterns against the state-of-art techniques.

IV. EXPERIMENTAL SETUP

The experimental work is carried out in JAVA language for evaluating the matching sequences pattern in chemical bonding. The performance of proposed technique is evaluated with parameters such as density of chemical bonds, size of chemical bond structural, classification accuracy, and number of matched sequential patterns against existing state-of-art techniques. The experimental data used for the analysis of proposed and existing techniques are extracted from Molecular Description Data Sets (Octane Isomers (O8), PolyAromatic Hydrocarbons (PAH), and PolyChloroBiphenyls (PCB)). The data set description is provided in Table 1.

Table 1 Molecular Description Data Set (Octane isomers)

S. No	Properties	Description
1	BP	Boiling Point
2	MP	Melting point
3	CT	Heat capacity at T constant
4	CP	Heat capacity at P constant
5	S	Entropy
6	DENS	Density
7	HVAPP	Enthalpy of vaporization
8	DHVAP	Standard enthalpy of vaporization
9	HFORM	Enthalpy of formation
10	DHFORM	Standard enthalpy of formation
11	MON	Motor Octane Number
12	MR	Molar Refraction
13	AcenFac	Accentric Factor
14	TSA	Total Surface Area
15	LogP	Octanol Water Coefficient
16	MV	Molar Volume

V. DISCUSSION

In this section we present the experiment results in terms of chemical bond density, chemical bond size, chemical bond classification time, chemical bond classification accuracy and number of matched sequential patterns. For the purpose of exploring the statistical significance of the results, we statistically compared two techniques, Probabilistic Sequence Translation Alignment Model (PSTAM) [1] and Fast Clustering based Feature Selection algorithm (FAST) [2].

A. Chemical bond classification time

Chemical bond classification time is measured using the number of chemical bonds (i.e. density) and the classification time to identify new bonds. The mathematical formulation for Chemical bond classification time is given as below.

$$CT = \sum_{d=1}^{n} CB_d * Time\left(Prob\left(Cl, Att_1, Att_2, \ldots, Att_n\right)\right) \quad (9)$$

From (9), the classification time 'CT' is measured using the number of chemical bonds or density of chemical bonds 'CB_d' and the time taken to perform the classification using probabilistic classifier and measured in terms of milliseconds. To measure the classification time, with chemical bond density of 15, chemical bond classification time using PDF-SPM was observed to be 5.25ms, 6.15ms using PSTAM and 6.6ms using FAST. The results are presented in table 2.

Table 2 Chemical bond classification time

Chemical bond density	Chemical bond classification time (ms)		
	PDC-SPM	PSTAM	FAST
15	5.29	6.17	6.63
30	8.34	10.14	11.24
45	13.14	15.34	16.44
60	18.19	20.43	21.53
75	23.14	25.13	26.23
90	28.24	30.34	31.44
105	32.14	34.64	35.64

To assess the performance of PDC-SPM technique and compare it to other techniques, namely, PSTAM [1] and FAST [2] all three techniques were implemented using JAVA with Weka tool and tabulation are shown in table 2.

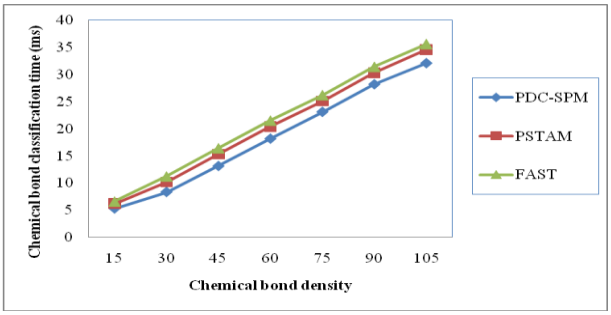

Figure 6 Measure of Chemical bond classification time

To estimate the chemical bond classification time, the density of chemical bonds are combined and the product of the density of chemical bonds is performed based on the samples taken from 18 Octane isomers. To measure chemical bond classification time, the melting point, boiling point, heat capacity at T constant and heat capacity at P constant were considered. Figure 6 illustrates the impact of changes in the chemical bond classification time on different sample periods (i.e. with different chemical bonds). As illustrated in Figure 6 the chemical bond classification time for chemical dataset c9 using different sample periods.

The chemical bond classification time is reduced using PDS-SPM technique by 13% compared to PSTAM [1] that helps in better classification of chemical bonds using the Probabilistic Deterministic Classifier. Moreover, by applying Probabilistic Deterministic Classifier in PDS-SPM technique applies Pearson's cumulative test that extracts the results based on the closeness of two atomic probabilities and therefore reduces the chemical bond classification time by 20% compared to FAST [2].

B. Classification accuracy

The classification accuracy is the measure to determine the rate of probability of the structures of chemical bonds being classified. The classification accuracy is mathematically formulated as given below.

$$CA = \sum_{i=1}^{n}((Chemical\ bonds\ properly\ classified)/CB_i) * 100 \quad (10)$$

From (10), the classification accuracy 'CA' is measured with chemical bonds properly classified to the total chemical bonds 'CB_i' used in the experimentation. It is measured in terms of percentage and higher the classification accuracy more efficient the method is said to be. Out of 30 chemical bond density, the chemical bond properly classified

using PDC-SPM was 27, 24 using PSTAM and 21 using FAST. Table 3 provides the chemical bond classification accuracy values obtained from (10).

Table 3 Chemical bond classification accuracy

Chemical bond density	Chemical bond classification accuracy (%)		
	PDC-SPM	PSTAM	FAST
15	85.23	77.32	68.23
30	91.35	82.14	72.28
45	93.47	84.15	75.89
60	86.17	75.78	68.32
75	88.24	78.32	80.14
90	92.15	82.14	74.69
105	94.32	83.12	75.90

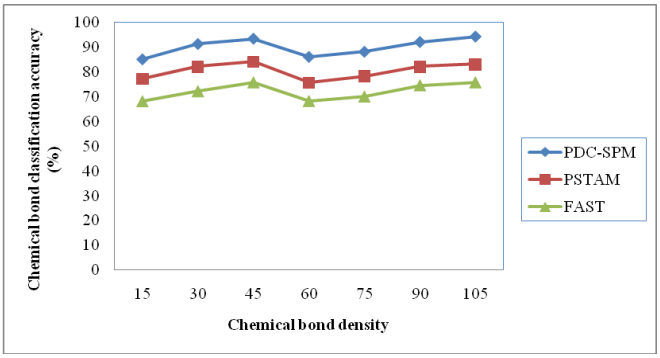

Figure 7 Measure of classification accuracy

Figure 7 compares the chemical bond classification accuracy for different samples (i.e. chemical bond density) using PDC-SPM technique to that of PSTAM and FAST for the similar scenarios discussed above. In all scenarios, the PDC-SPM technique outperforms all two systems. As illustrated in the graphs, the chemical bond classification accuracy is not linear which states that the association of atoms into molecules, ions, crystals, and other stable species that make up the familiar substances in chemical information differ with different chemical bond size.

It can be observed that with the chemical bond density from 15 – 45, the chemical bond classification accuracy observed using all three techniques increased whereas with chemical bond density in the range of 45 – 60, the chemical bond classification accuracy decreased using PDC-SPM technique in comparison to two other techniques [1] and [2]. This is because of the application of Naive Bayes Probabilistic Deterministic Classifier algorithm. Using the Pearson's cumulative test in Naive Bayes Probabilistic Deterministic Classifier algorithm, set of two probabilities with total valence electrons disseminated between two atoms resulting in maximizing the chemical bond classification accuracy. As a result, the maximum chemical bond classification accuracy increases by 11% compared to PSTAM and 20% compared to FAST respectively.

C. Number of matched sequential patterns

The number of matched sequential patterns 'MSP' measures the effectiveness of the technique. It is the difference between the total sequential patterns generated 'TSP' and incorrect sequential patterns 'ISP' and is expressed as given below.

$$MSP = TSP - ISP \quad (11)$$

From (11), higher the number of matched sequential patterns, more efficient the technique is said to be. The chemical bond size is measured in terms of picometre 'pm'. Table 4 shows the results of matched sequential patterns obtained using the three techniques, namely, PDC-SPM, PSTAM and FAST respectively.

Table 4 Number of matched sequential patterns

Chemical bond size (pm)	Number of matched sequential patterns		
	PDC-SPM	PSTAM	FAST
25	14	12	9
50	22	19	16
75	35	31	28
100	51	48	41
125	65	60	55
150	78	71	67
175	91	85	80

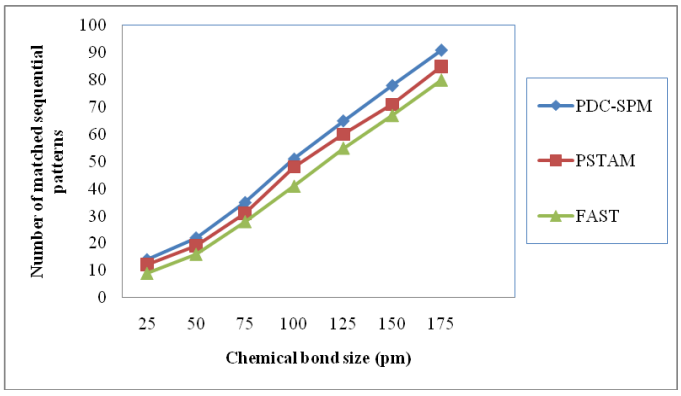

Figure 8 Measure of number of matched sequential patterns

Figure 8 shows the behaviour of the number of matched sequential patterns in response to total sequential patterns generated by varying the chemical bond size. The average number of matched sequential patterns of the three methods was observed to be increasing with the increasing chemical bond size in the range of 25pm and 175 pm. Despite equivalency, the number of matched sequential patterns was observed to be greater using the PDC-SPM technique when compared to PSTAM [1] and FAST [2] respectively. This is because of applying Sequential Pattern Chemical Structure Bond Mining (SPCSBM) algorithm that initially measures the bond, then constructs the equivalent graphs and through which similarity between vertices based on maximal chemical bond. This in turn improves the matching rate using PDC-SPM by 10% compared to PSTAM. Besides, using maximum chemical bond function, maximum pair-wise chemical bond similarity between molecules are significantly calculated with the aid of SPCSBM algorithm form an improvement in the number of matched sequential patterns by 21% compared to FAST.

VII. CONCLUSION

In this work, an effective technique called Probabilistic Deterministic Classifier and Sequential Pattern Mining (PDC-SPM) is presented. The technique improves the number of matched sequential patterns that in turn reduces the chemical bond classification time based on the chemical bond size and density. The goal of PDC-SPM is to improve the chemical bond classification accuracy with the identified structure of chemical bonds as belonging to a specific class and therefore to evaluate the structural patterns of chemical bonding which significantly contribute to the relevance. To do this, we first designed a Probabilistic Deterministic Classifier model that measures the probability of membership in each class of chemical bonds based on the probability function. Then, based on this measure, a Sequential Pattern Chemical Structure Bond Mining (SPCSBM) algorithm is designed for identifying different sequential patterns in an extensive manner. Extensive experiments were carried out and compared with the state-of-the art methods. The results show that PDC-SPM technique offers better performance with an improvement of chemical bond classification accuracy by 15% and reduces the time taken for chemical bond classification by 16% compared to PSTAM and FAST respectively.

REFERENCE

[1] Minyoung Kim, "*Probabilistic Sequence Translation-Alignment Model for Time-Series Classification*", IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issue 2, February 2014, Pages 426-437.

[2] 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, Volume 25, Issue 1, January 2013, Pages 1-14.

[3] T.Ramraj, R.Prabhakar, "*Frequent Subgraph Mining Algorithms –A Survey*", Elsevier, Procedia Computer Science 47 (2015) 197 – 204.

[4] David Gotz , Fei Wang, Adam Perer, "*A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data*", Elsevier, Journal of Biomedical Informatics, Volume 48, April 2014, Pages 148–159.

[5] Igor Chikalov, Peggy Yao, Mikhail Moshkov, Jean-Claude Latombe, "*Learning probabilistic models of hydrogen bond stability from molecular dynamics simulation trajectories*", Springer, BMC Bioinformatics. 2011; 12(Suppl 1): S34, February 2011, Pages 1-6.

[6] Chunlin Hao., Sheng Huang., Zixin Deng, Changming Zhao, Yi Yu, "*Mining of the Pyrrolamide Antibiotics Analogs in Streptomyces netropsis Reveals the Amidohydrolase- Dependent ''Iterative Strategy'' Underlying the Pyrrole Polymerization*", Plos One, Volume 9, Issue 6, June 2014, Pages 1-9.

[7] Shubhra Sankar Ray and Sankar K. Pal, "*RNA Secondary Structure Prediction Using Soft Computing*", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume 10, Issue 1, January/February 2013, Pages 2-17.

[8] Osvaldo Navarro, Rene Cumplido, Luis Villasen or-Pineda, Claudia Feregrino-Uribe, Jesu´ s Ariel Carrasco-Ochoa, "*A Node Linkage Approach for Sequential Pattern Mining*, Plos One, Volume 9, Issue 6, June 2014, Pages 1-16.

[9] Ron Henkel, Fabienne Lambusch, Olaf Wolkenhauer, Kurt Sandkuhl, Christian Rosenke and Dagmar Waltemath, "*Finding Patterns in Biochemical Reaction Networks*", peerj.preprints.1479v2, June 2016, Pages 1-12.

[10] Shinya Suzumura, Kazuya Nakagawa, "*Selective Inference Approach for Statistically Sound Discriminative Pattern Discovery*", February 2016, Pages 1-28.

[11] Hyeon Ah Park , Taewook Kim , Meijing Li , Ho Sun Shon ,Jeong Seok Park , Keun Ho Ryu, "*Application of Gap-Constraints Given Sequential Frequent Pattern*

Mining for Protein Function Prediction", Elsevier, Osong Public Health and Research Perspectives, Volume 6, Issue 2, April 2015, Pages 112–120.

[12] Hao Yang, Lei Duan, Guozhu Dong, Jyrki Nummenmaa, Changjie Tang, and Xiaosong Li, "*Mining Itemset-Based Distinguishing Sequential Patterns with Gap Constraint*", Springer, Database Systems for Advanced Applications, Volume 9049, April 2015, Pages 39-54.

[13] Jerry Chun-Wei Lin1, Wensheng Gan1, And Tzung-Pei Hong, "*Efficiently Maintaining the Fast Updated Sequential Pattern Trees With Sequence Deletion*", IEEE Access , Volume 2, Issue 2014, November 2014, Pages 1374 – 1383.

[14] Claude Pasquier, Jérémy Sanhes, Frédéric Flouvat, Nazha Selmaoui-Folcher, "*Frequent pattern mining in attributed trees: algorithms and applications"*, Springer, Knowledge and Information Systems, March 2016, Volume 46, Issue 3, Pages 491-514.

[15] Fabio Fumarola, Pasqua Fabiana Lanotte, Michelangelo Ceci,Donato Malerba, "*CloFAST: closed sequential pattern mining using sparse and vertical id-lists*", Springer, Knowledge and Information Systems, Pages 1-35.

[16] Aileen P. Wright , Adam T. Wright b, Allison B. McCoy c, Dean F. Sittig, "*The use of sequential pattern mining to predict next prescribed medications*", Elsevier, Journal of Biomedical Informatics, Volume 53, February 2015, Pages 73–80.

[17] Wenjing Zhang and Xin Feng, "*Event Characterization and Prediction Based on Temporal Patterns in Dynamic Data System*", IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issue 1, January 2014, Pages 144-156.

[18] Massimiliano Pierobon, and Ian F. Akyildiz, "*A Statistical–Physical Model of Interference in Diffusion-Based Molecular Nanonetworks*", IEEE Transactions on Communications, Volume 62, Issue 6, June 2014, Pages 2085-2095.

[19] Peggy Cellier, Thierry Charnois, Marc Plantevit, Christophe Rigotti, Bruno Crémilleux, Olivier Gandrillon, Jiˇrí Kléma and Jean-Luc Manguin, "*Sequential pattern mining for discovering gene interactions and their contextual information from biomedical texts*", Cellier et al. Journal of Biomedical Semantics (2015) 6:27, June 2015, Pages 1-12.