Reducing Square-Error of Jarvis-Patrick Algorithm for Drug Discovery

Ashraf B. El-Sisi, Hamdy M. Mousa, Mohamed G. Malhat
Computer Science dept., Faculty of Computers and Information,
Menofia University, Egypt

***Abstract*—Clustering algorithms play an important role in chemoinformatics and especially in the drug discovery process. Clustering methods may be hierarchical or non-hierarchical. Non-hierarchical algorithms have fast processing for clustering large chemical data sets than hierarchical algorithms. One of the most popular non-hierarchical clustering algorithms that are used in many applications in the drug discovery process is Jarvis-Patrick algorithm. The applications of Jarvis-Patrick in the drug discovery process are compound selection, compound acquisition, low-throughput screening and Quantitative Structure-Activity Relationship (QSAR) analysis. Jarvis-Patrick groups compounds in a cluster based on a three neighborhood conditions. These three conditions groups compounds, which are not similar enough, in the same cluster. Adding dissimilar compounds in the same cluster will lead to poor compound selection, compound acquisition and QSAR analysis. In this paper, standard Jarvis-Patrick is modified by adding a fourth condition which computed only if the three standard conditions are true. This condition computes the increasing in the value of Square Error (SE) of the cluster by adding a compound and compares it with expected increasing in SE to determine whether to add a compound to the cluster or not. The result shows that our modification produces clusters with more similar compounds and still has fast processing.**

Keywords—Chemoinformatics; Drug Discovery; Non-hierarchical Clustering; Jarvis-Patrick

I. INTRODUCTION

The use of clustering for chemical applications is based on similar property and activity principle which states that compounds with similar structures are likely to exhibit similar properties, which known as Structure-Property Relationship (SPR), and similar activities which known as Structure-Activity Relationship (SAR) [1]. Clustering algorithms, which are used in chemical application, must group more similar compounds in term of properties or activity in the same cluster based on their structure. Most clustering algorithms for chemical application cover the area of drug discovery process [2, 3]. The drug discovery is the process of making drugs that response to diseases with fewer side effects. It consists of seven steps: disease selection, target hypothesis, leads compound identification, lead optimization, pre-clinical trial, and clinical trial and pharmacogenomic optimization [4].

Chemoinformatics are used in lead compound identification and optimization steps [5]. Chemoinformatics are the application of informatics methods that are used to solve chemical problems. It is a new discipline emerging from storing, manipulating, processing, design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information. The use of chemoinformatics becomes a critical part of the drug discovery process as it accelerates the drug discovery process and reduces the overall cost [6, 7]. There are many applications of chemoinformatics in the drug discovery such as compound selection, compound acquisition, virtual library generation, virtual screening, QSAR analysis and Absorption, Distribution, Metabolism, Elimination, and Toxicity (ADMET) prediction [8-11]. Central tasks of most of these applications are the establishment of a relationship between a chemical structure and its biological activity and the prediction of pharmacological properties in addition to lead finding [5, 6].

Clustering algorithms are used in most of these applications as a method of selection, diversity analysis and data reduction. Compared to the other costs of drug discovery, clustering can add significant value at minimal cost [12]. Clustering algorithms divided into two main categories hierarchical and non-hierarchical. Jarvis-Patrick is one of the most popular non-hierarchical clustering algorithms that has a wide range of applications in chemoinformatics because of it is fast processing for clustering large chemical data sets and ease implementation. Standard implementation of Jarvis-Patrick may group compounds in one cluster that are not similar enough because the compounds satisfy the three neighborhood conditions. Adding dissimilar compounds in the same cluster will increase the value of SE in clusters and lead to increase in the SSE (the sum of SE for all clusters) of the produced clusters. SSE is one of the quality measures that used to evaluate clustering algorithm in its ability to group more similar compounds in the same cluster.

Standard Jarvis-Patrick is modified by adding a condition that will be computed only if the standard Jarvis-Patrick conditions are true. This condition will determine if to add a compound to a cluster or not. The condition computes the increasing in value of Square Error (SE) of the cluster by adding this compound and compares it with expected increasing in SE. If this increasing is less than or equal to the expected increasing then the compound will be added to the cluster else the compound will not be added. The results show that by adding this condition, Jarvis-Patrick will not add dissimilar compounds to the same cluster and still has fast processing. The organization of this paper is as following. In section 2, standard Jarvis-Patrick and its usage in chemoinformatics are overviewed. In section 3, our modification on Jarvis-Patrick is proposed. In section 4, modified Jarvis-Patrick is compared with standard Jarvis-Patrick and their implementation and experimental results are discussed. Finally in section 5, conclusion is given.

II. JARVIS-PATRICK CLUSTERING USAGE IN CHEMOINFORMATICS

Clustering methods are used in a number of disciplines such as computer science, information technology, information system, engineering, bioinformatics and chemoinformatics. The main using of clustering methods in chemoinformatics is to group similar compounds in a cluster based on the underlying distribution of input. After grouping these compounds, the activity of compound is predicted based on known compounds activity that are in the same cluster.

Jarvis-Patrick is one of the most popular methods that have a wide range of applications in chemoinformatics because of its ability to handle large data sets in reasonable time, ease implementation and the availability of an efficient commercial implementation from Daylight for handling very large data sets [13]. Jarvis-Patrick is non-hierarchical non-overlapping clustering method. Non-overlapping means that each compound can be only in one cluster. Non-hierarchical means that data set is analyzed to produce a single partition of the compounds resulting in a set of clusters.

Standard Jarvis-Patrick method proceeds in two levels [14]. In the first level, a list of the top K nearest neighbors (K is usually16) is generated for each compound in the data set. The nearest neighbors are usually determined by the Euclidean distance for numerical descriptor and by the Tanimoto coefficient for binary descriptor [15]. In the second level, the nearest-neighbor lists are scanned to create clusters that satisfy the three following neighborhood conditions:

1. The top K nearest-neighbor list of compound i must contain compound j.
2. The top K nearest-neighbor list of compound j must contain compound i.
3. The top K nearest-neighbor lists of compound i and j must have at least K-Min common compounds (Kmin is determined by user and in the range 1 to K).

The pairs of compounds, that don't satisfy any of the above three conditions, are not put into the same cluster. The value of top K nearest-neighbors specifies the number of compound's neighbors to consider when counting the number of mutual neighbors shared with another compound. This value must be at least 2. Lower values make the algorithm to finish faster, but the final set of clusters will have many small clusters. Higher values cause the algorithm to take longer time to finish, but may result in fewer clusters and clusters that form longer chains. The K-Min specifies the minimum number of mutual nearest neighbors that the two compounds must have to be in the same cluster. This value must be at least 1 and must not exceed the value of the K nearest-neighbors. Lower values result in clusters that are compact. Higher values result in clusters that are more dispersed.

The standard implementation of Jarvis-Patrick produces a large number of singletons and clusters with large SSE. Several modifications have been developed to overcome singletons problem such as:

1. A variable-length nearest-neighbor list [16], a proximity threshold is used to determine a variable number of neighbors for each compound. All neighbors that pass the threshold test are considered as neighbors to this compound. By this modification, outliers are prevented from joining a cluster while preventing the arbitrary splitting of large clusters arising from the limitations imposed by fixed length lists.

2. Re-clustering of singletons [17], standard Jarvis–Patrick is applied in an iterative way to remove the singletons. The singletons are assigned to a cluster using less strict parameters than defined by user. This iterative way is repeated until a fewer a specified percentage of singletons remain.

3. Fuzzy clustering [18], all compounds are assigned a probability that determines the distances of compounds from each cluster. The singletons are assigned to its nearest cluster based on specified threshold probability. For singletons that not exceed threshold, they will be regarded as outliers and remains as singletons.

The applications of Jarvis–Patrick clustering in chemoinformatics are compound selection, compound acquisition and high throughput screening. In [19], Jarvis-Patrick is used to cluster a data set of about 240,000 compounds for compound selection. Singletons are moved to the nearest non-singleton cluster. Then cluster centroids are calculated for each cluster to select representative compounds based on their closet centroid. In [20], Jarvis Patrick is to assist low-throughput screening and to support QSAR analysis by analyzing databases for efficient compound acquisition. In [17], Jarvis–Patrick is used for high throughput screening by the selection of compounds from the corporate database. In [18], Jarvis-Patrick is used for analysis of the compound database to support high throughput screening.

The previous modifications are developed to overcome the singletons problem. The three neighborhood conditions of Jarvis-Patrick don't guarantee to group more similar compounds in the same cluster. So, the produced clusters have large SSE values. In the next section, the standard Jarvis-Patrick algorithm will be modified by adding a fourth condition to overcome this problem.

III. PROPOSED MODIFICATION ON STANDARD JARVIS-PATRICK

The standard Jarvis-Patrick will be modified by adding a fourth neighborhood condition that will be computed only if the three previous neighborhood conditions are true. The fourth condition will compute the increasing in SE for a cluster contains compound i after adding compound j to this cluster and compare it with expected increasing in SE. First, for the cluster of n compounds each represented by a vector. The vector of the cluster centroid, x(c), is defined as

$$\mathrm{X(c)} = \left({}^{1}\!/_{n}\right) \sum_{r=1}^{n} \mathrm{x(r)} \qquad (1)$$

The centroid is the simple arithmetic mean of the vectors of the cluster members. The SE for a cluster is the sum of squared Euclidean distances to the centroid for all n compounds in that cluster. The SE is defined as

$$\text{SE} = \sum_{r=1}^{n}[X(r) - X(c)]^2 \tag{2}$$

The SSE is the summation of SE for all produced m clusters and is defined as

$$\text{SSE} = \sum_{s=1}^{m} SE_s \tag{3}$$

The increasing in SE is the difference between the value of SE for the cluster containing i after adding compound j and before adding compound j. The increasing in SE is defined as

$$\text{Increasing in SE} = SE_{after\ adding\ j} - SE_{before\ adding\ j} \tag{4}$$

The expected increasing in SE is the SE for data set divided by number of compounds n multiplied with a user specified ratio r; r is a value between 0 and 1. Small values of r will ensure that more similar compounds will be grouped into the same cluster. The expected increasing in SE is defined as

$$\text{Expected increasing in SE} = \frac{SE_{data\ set}}{n} * r \tag{5}$$

If increasing in SE is less than or equal expected increasing, then compound j will be added to the cluster containing compound i, else compound j will not be added to this cluster. By adding this modification, fourth condition will produce clusters with less SSE by not adding the compounds that will increase SE than expected increasing into the same cluster. So, compound selection, acquisition and QSAR analysis will be more efficient and the algorithm still has fast processing because the fourth condition will not be computed only if the three conditions of standard Jarvis-Patrick algorithm are true.

IV. Implementation nd Experimental Results

The implementations of the algorithms are in JAVA, under Windows-7 operating system, Intel core-i5, 2.5 GHz and Ram 4 GB. NCI data set, one of the most popular data set, is used for experimental [21]. Three random subsets are taken from NCI data set with the following number of compounds and SE as shown in Table 1.

TABLE I. THREE SUBSETS OF NCI DATA SET

Subset Name	Number of Compounds	SE
NCI-1	100	25.61546473
NCI-2	500	791.56501
NCI-3	1000	1838.0002

BCUT descriptor is used to represent compounds in the three subsets [22]. For each NCI subset, 4 runs are recorded with K=16 and K-Min= 4, 8, 12 and 14 for each run. Table 2 shows the K, K-Min, Number of Clusters (NOC), Computation time in milliseconds and SSE of standard Jarvis Patrick algorithm. Tables 3, 4, 5 and 6 show the same information for modified Jarvis Patrick algorithm where r = 1.0, 0.5, 0.1 and 0.01.

TABLE II. OUTPUT OF STANDARD JARVIS-PATRICK ALGORITHM

Data set Name	K	K-Min	NOC	SSE	Time in Milliseconds
NCI-1	16	4	8	13.29864	40
	16	8	10	4.069484	20
	16	12	28	1.49243	10
	16	14	62	0.238635	10
NCI-2	16	4	46	44.92072	190
	16	8	63	29.30268	140
	16	12	200	14.63466	130
	16	14	335	2.768861	120
NCI-3	16	4	85	43.46274	480
	16	8	126	28.03772	420
	16	12	387	11.34654	410
	16	14	683	6.63129	410

TABLE III. OUTPUT OF MODIFIED JARVIS-PATRICK ALGORITHM WHERE R = 1.0

Subset Name	K	K-Min	NOC	SSE	Time in Milliseconds
NCI-1	16	4	10	4.926679	80
	16	8	11	3.459389	30
	16	12	28	1.457589	30
	16	14	62	0.232923	10
NCI-2	16	4	48	39.88919	270
	16	8	65	24.27115	150
	16	12	201	11.79896	140
	16	14	335	2.768861	140
NCI-3	16	4	84	40.29349	620
	16	8	126	23.87385	460
	16	12	378	11.19564	450
	16	14	659	6.579031	430

TABLE IV. OUTPUT OF MODIFIED JARVIS-PATRICK ALGORITHM WHERE R = 0.5

Subset Name	K	K-Min	NOC	SSE	Time in Milliseconds
NCI-1	16	4	13	2.894775	90
	16	8	13	2.878076	30
	16	12	30	0.858161	20
	16	14	62	0.232923	20
NCI-2	16	4	52	29.29845	270
	16	8	67	16.73628	180
	16	12	203	7.632148	150
	16	14	336	2.705852	140
NCI-3	16	4	84	39.44875	610
	16	8	126	23.02911	460
	16	12	378	9.807962	430
	16	14	659	5.191356	410

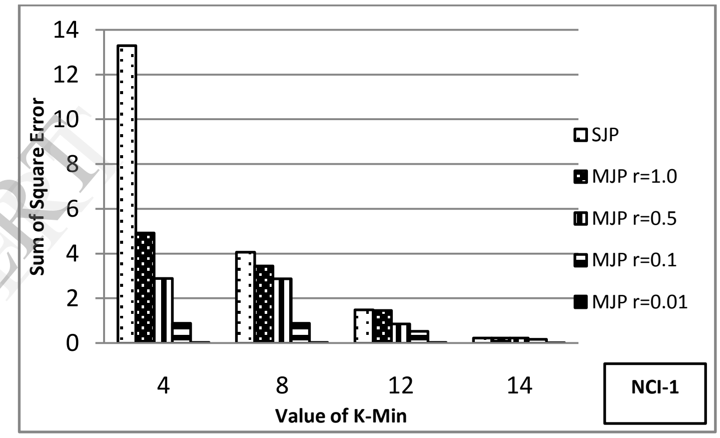
TABLE V. OUTPUT OF MODIFIED JARVIS-PATRICK ALGORITHM WHERE R = 0.1

Subset Name	K	K-Min	NOC	SSE	Time in Milliseconds
NCI-1	16	4	25	0.897523	70
	16	8	25	0.896463	50
	16	12	36	0.530665	20
	16	14	64	0.177409	10
NCI-2	16	4	61	11.57974	290
	16	8	76	8.472676	170
	16	12	208	3.491306	150
	16	14	336	1.312308	140
NCI-3	16	4	91	25.03054	600
	16	8	133	16.17223	480
	16	12	382	5.916373	440
	16	14	662	2.095064	440

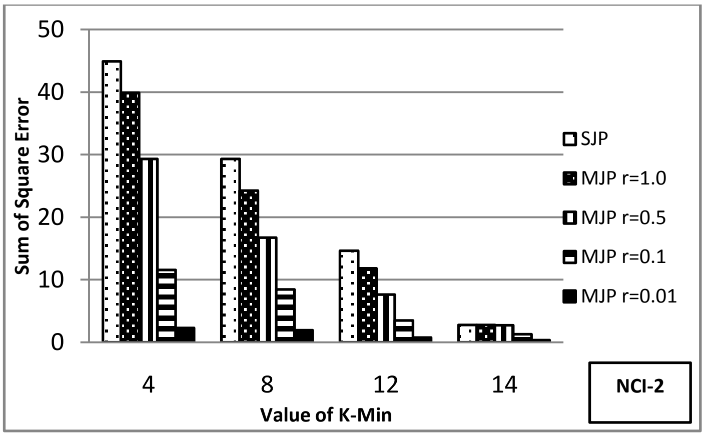
TABLE VI. OUTPUT OF MODIFIED JARVIS-PATRICK ALGORITHM WHERE R = 0.01

Subset Name	K	K-Min	NOC	SSE	Time in Milliseconds
NCI-1	16	4	70	0.031339	70
	16	8	70	0.031339	50
	16	12	72	0.02826	30
	16	14	78	0.01946	10
NCI-2	16	4	117	2.293131	270
	16	8	130	1.933292	180
	16	12	237	0.764977	140
	16	14	348	0.368605	140
NCI-3	16	4	171	5.603048	570
	16	8	198	4.629283	470
	16	12	420	1.893126	440
	16	14	677	0.592186	430

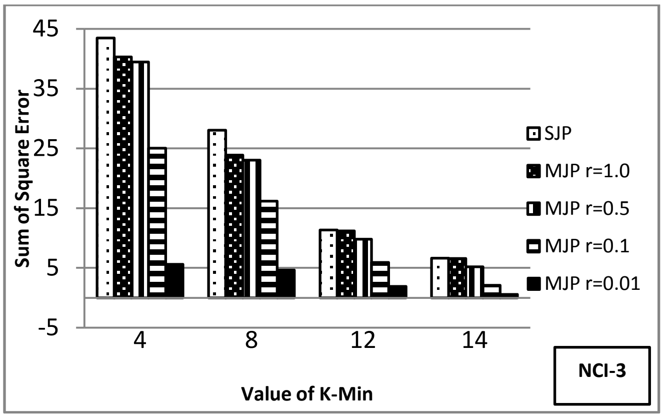
Fig.1 shows the SSE for the Standard Jarvis-Patrick (SJP) and Modified Jarvis-Patrick (MJP) where r = 1.0, 0.5, 0.1 and 0.01 for the three subsets. As shown in Fig.1, our approach produces clusters with less or equal SSE than SJP for all subsets with K-Min = 4, 8, 12 and 14. For example in NCI-1 subset when K-Min = 4, SJP produces clusters with SSE equal 13.2986 and MJP produces clusters with SSE equal 4.9266 where r = 1.0, 2.8947 where r = 0.5, 0.8975 where r = 0.1 and 0.0313 where r = 0.01. When K-Min = 14, SJP produces clusters with SSE equal 0.2386 and MJP produces clusters with SSE equal 0.2329 where r = 1.0, 0.2329 where r = 0.5, 0.1774 where r = 0.1 and 0.0194 where r = 0.01. From previous results, as the value of K-Min increase, MJP produces clusters with SSE less than or equal to SJP. When K-Min decrease, MJP produces clusters with SSE less than SJP for all values of r.



(a)



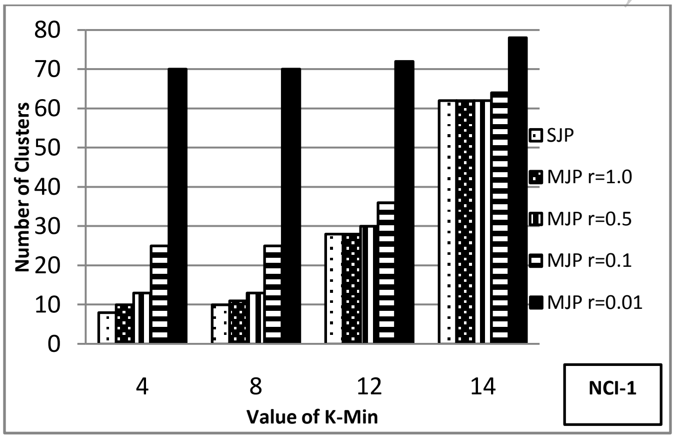
(b)



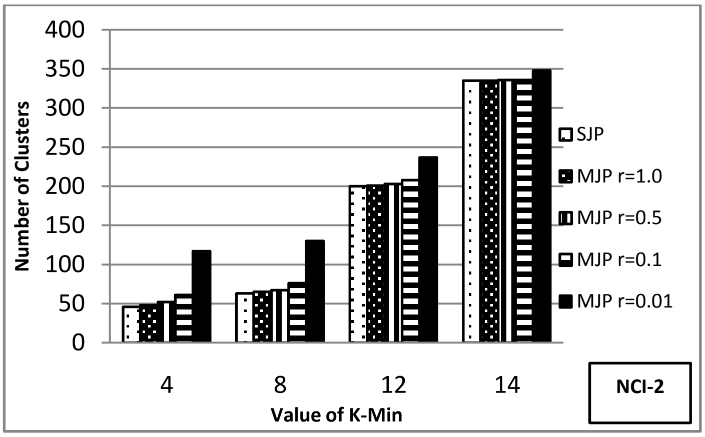
(c)

Figure 1. SSE of SJP and MJP for three subsets where r = 1.0, 0.5, 0.1 and 0.01

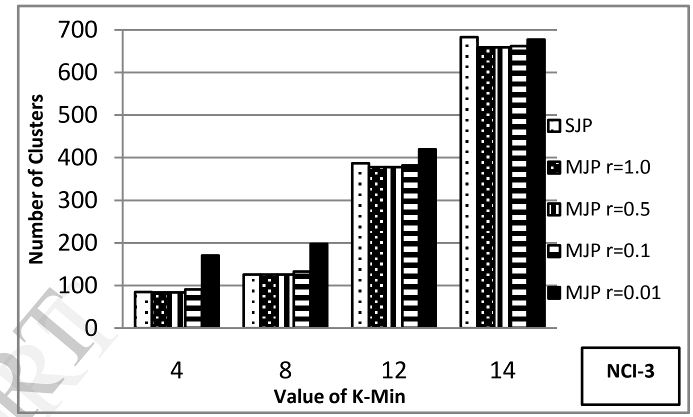
Fig.2 shows the number of clusters generated by SJP and MJP where r = 1.0, 0.5, 0.1 and 0.01 for the three subsets. As shown in Fig.2, the number of clusters generated by our approach is large than or equal to the number of clusters generated by SJP for all subsets with K-Min = 4, 8, 12 and 14. For example in NCI-1 subset when K-Min = 4, SJP produces 8 clusters and MJP produces 10 clusters where r = 1.0, 13 clusters where r = 0.5, 25 clusters where r = 0.1 and 70 clusters where r = 0.01. When K-Min = 14, SJP produces 62 clusters and MJP produces 62 clusters where r = 1.0, 62 clusters where r = 0.5, 64 clusters where r = 0.1 and 78 clusters where r = 0.01. From previous results, as the value of K-Min increase MJP and SJP produce similar number of clusters and when K-Min decrease MJP produces more clusters than SJP for all values of r.



(a)



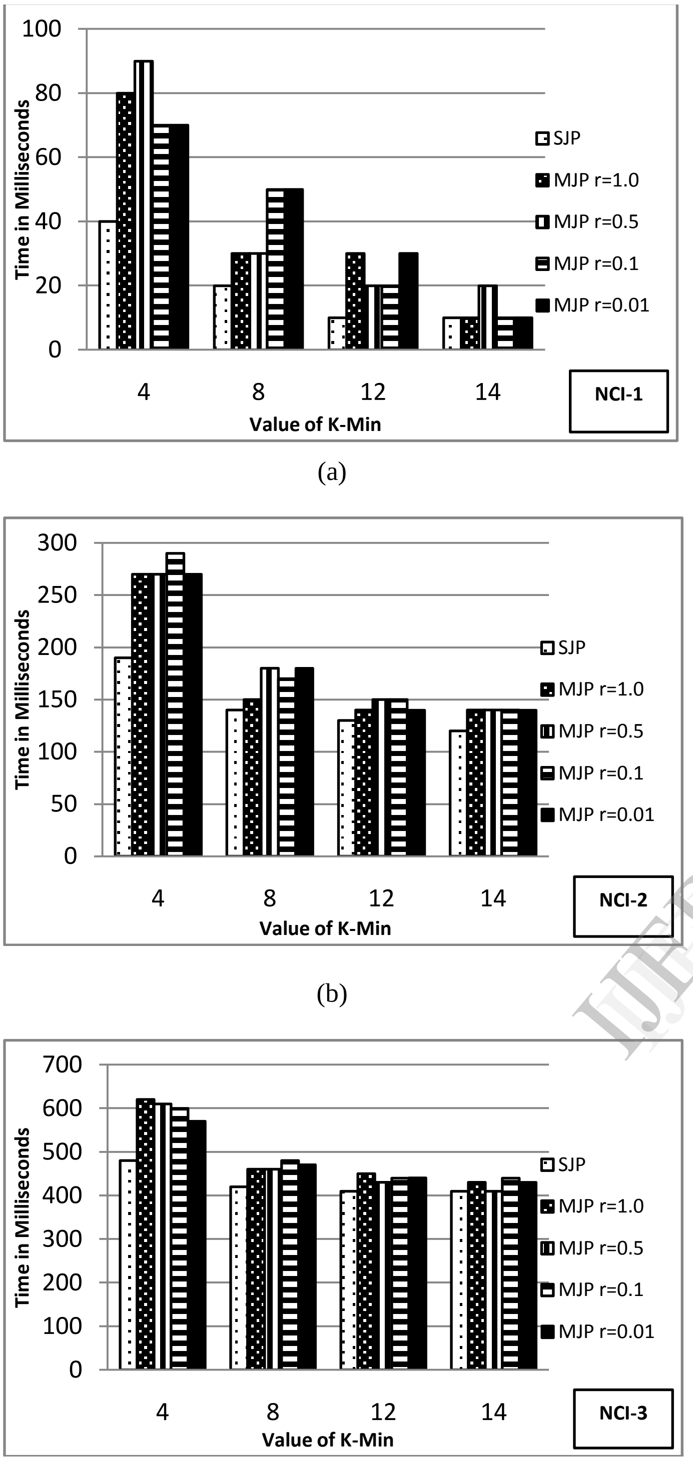
(b)



(c)

Figure 2. Number of Clusters of SJP and MJP for three subsets where r = 1.0, 0.5, 0.1 and 0.01

Fig.3 shows the time required in milliseconds for SJP and MJP where r = 1.0, 0.5, 0.1 and 0.01 for the three subsets. As shown in Fig.3, The time required for our approach is large than or equal to the time required for SJP for all subsets with K-Min = 4, 8, 12 and 14. For example in NCI-1 subset when K-Min = 4, SJP takes 60 milliseconds and MJP takes 60 milliseconds where r = 1.0, 90 milliseconds where r = 0.5, 70 milliseconds where r = 0.1 and 70 milliseconds where r = 0.01. When K-Min = 14, SJP takes 10 milliseconds and MJP takes 10 milliseconds where r = 1.0, 20 milliseconds where r = 0.5, 10 milliseconds where r = 0.1 and 10 milliseconds where r = 0.01. From previous results, as the value of K-Min increase MJP and SJP take similar computation time and when K-Min decrease MJP takes more time than SJP for all values of r. The increasing in time for MJP represents the overhead time needed to process the fourth condition.



(a)

(b)

(c)

Figure 3. Time Required for SJP and MJP for three subsets where r = 1.0, 0.5, 0.1 and 0.01

Form Figures 1-3, our approach results reduce SSE in the resulted clusters than SJP. This SSE reducing is obvious for small values of K-Min because small values of K-Min will give the opportunity for the fourth condition to be invoked and the percentage of SEE reducing is depending on value of r. If the value of r is small then the SSE is more reduced. In order to reduce SSE, more clusters will be generated. These extra clusters represent compounds that don't satisfy the fourth condition. The increasing in time needed by our approach is overhead time to apply the fourth condition.

V. CONCLUSION

The demands of clustering data sets of several million compounds with high-dimensional representations led to the widespread adoption of a few inherently efficient and optimally implemented methods. Jarvis-Patrick is one of the most popular clustering methods that have many applications in chemoinformatics such as compound selection, compound acquisition, lead-finding and QSAR analysis. In this paper, standard Jarvis-Patrick is modified in order to group more similar compounds in the same cluster and avoiding adding compounds to clusters that will increase SSE. The results show that our modification produces clusters with less SSE than standard Jarvis-Patrick. So, compound selection, acquisition and QSAR analysis will exhibit better efficiency and at the same time Jarvis-Patrick still has fast processing. In the future work, Modified Jarvis-Patrick will be applied for large chemical data sets and will be compared with ward clustering algorithm.

REFERENCES

[1] Geoffrey M. Downs and Peter Willett, "The Use of Similarity and Clustering Techniques for the Prediction of Molecular Properties," in Applied Multivariate Analysis in SAR and Environmental Studies, J. Devillers and W. Karcher, Eds.: Springer Netherlands, 1991, vol. 2, pp. 247-279.

[2] J. Nouwen and B. Hansen, "An Investigation of Clustering as a Tool in Quantitative Structure-Activity Relationships (QSARS)," SAR and QSAR in Environmental Research, vol. 4, no. 1, pp. 1-10, 1995, PMID: 22091841.

[3] Kenny B. Lipkowitz and Donald B. Boyd, Reviews in Computational Chemistry. New York, NY, USA: John Wiley ; Sons, Inc., 2002.

[4] Thomas Engel, "Basic Overview of Chemoinformatics," Journal of Chemical Information and Modeling, vol. 46, no. 6, pp. 2267-2277, 2006, PMID: 17125169.

[5] Gyorgy M. Keseru and Gergely M. Makara, "Hit discovery and hit-to-lead approaches," Drug Discovery Today , vol. 11, no. 15, pp. 741-748, 2006.

[6] Charu C. Aggarwal and Haixun Wang, Managing and Mining Graph Data, 1st ed.: Springer Publishing Company, Incorporated, 2010.

[7] Andrew R. Leach and Valerie J. Gillet, An Introduction to Chemoinformatics.: Springer Publishing Company, Incorporated, 2007.

[8] Jeremy L. Jenkins, Andreas Bender, and John W. Davies, "In silico target fishing: Predicting biological targets from chemical structure ," Drug Discovery Today: Technologies , vol. 3, no. 4, pp. 413-421, 2006.

[9] Meenakshi Mishra, Hongliang Fei, and Jun Huan, "Computational Prediction of Toxicity," Int. J. Data Min. Bioinformatics, vol. 8, no. 3, pp. 338-348, 2013.

[10] Christian Korn and Stefan Balbach, "Compound selection for development - Is salt formation the ultimate answer? Experiences with an extended concept of the "100 mg approach"," European Journal of Pharmaceutical Sciences , vol. 57, no. 0, pp. 257-263, 2014, Special Issue on 7th International Symposium on Microdialysis - Edited By: William Couet and Hartmut Derendorf.

[11] Vishnu J. Gaikwad, "Application of Chemoinformatics for Innovative Drug Discovery," International Journal of Chemical Sciences and Applications, vol. 1, no. 1, pp. 16-24, 2010.

[12] S Kavi Priya and M Lingaraj, "Performance analysis of data clustering in rapid mediccal development," International Journal of Engineering Research and Science \& Technology, vol. 2, no. 2, pp. 115-122, 2013.

[13] Daylight. [Online]. http://www.daylight.com/

[14] R. A. Jarvis and Edward A. Patrick, "Clustering Using a Similarity Measure Based on Shared Near Neighbors," Computers, IEEE Transactions on, vol. C-22, no. 11, pp. 1025-1034, 1973.

[15] Peter Willett, John M. Barnard, and Geoffrey M. Downs, "Chemical Similarity Searching," Journal of Chemical Information and Computer Sciences, vol. 38, no. 6, pp. 983-996, 1998.

[16] Robert D. Brown and Yvonne C. Martin, "Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection," Journal of Chemical Information and Computer Sciences, vol. 36, no. 3, pp. 572-584, 1996.

[17] Paul R. Menard, Richard A. Lewis, and Jonathan S. Mason, "Rational Screening Set Design and Compound Selection: Cascaded Clustering," Journal of Chemical Information and Computer Sciences, vol. 38, no. 3, pp. 497-505, 1998.

[18] Thompson N. Doman, John M. Cibulskis, Michael J. Cibulskis, Patrick Dale McCray, and Dale P. Spangler, "Algorithm5: A Technique for Fuzzy Similarity Clustering of Chemical Inventories," Journal of Chemical Information and Computer Sciences, vol. 36, no. 6, pp. 1195-1204, 1996.

[19] Peter Willett, Vivienne Winterman, and David Bawden, "Implementation of nonhierarchic cluster analysis methods in chemical information systems: selection of compounds for biological testing and clustering of substructure search output," Journal of Chemical Information and Computer Sciences, vol. 26, no. 3, pp. 109-118, 1986.

[20] Malcolm J. McGregor and Peter V. Pallai, "Clustering of Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors," Journal of Chemical Information and Computer Sciences, vol. 37, no. 3, pp. 443-448, 1997.

[21] NCI Data set. [Online]. http://cactus.nci.nih.gov/download/nci/.

[22] BCUT Descriptor. [Online]. http://sourceforge.net/projects/cdk/.