ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20151008153755

SELF-ORGANIZING MAPS WITH SLIDING WINDOW (SOM+SW)

Ulaş Çelenk, Duygu Çelik Ertuğrul, Metin Zontul, Osman Nuri Uçan

Original scientific paper
SOM is a popular artificial neural network algorithm to perform rational clustering on many different data sets. There is a disadvantage of the SOM that can run on a predefined completed data set. Various problems are encountered on a time-stream data sets when clustering by using standard SOM since the time-stream data sets are generated dependent on time. In this study, the Sliding Window feature is included into standard SOM for clustering time-stream data sets. Thus, the combination of SOM and Sliding Window (SOM + SW) gives more accurate results when clustering on time-stream data sets. To prove this, a set of internet usage data from a mobile operator in Turkey is taken to test. The taken data set from the mobile operator is clustered according to the classical SOM then the future data usages of subscribers are estimated. The same data set is applied on the SOM + SW to perform the same simulations.

Keywords: *clustering; mobile operators; self-organizing maps (SOM); sliding window; time-stream data sets*

Samoorganizirane mape s kliznim prozorom (SOM + SW)

Izvorni znanstveni članak
SOM je popularan algoritam umjetne neuronske mreže za obavljanje racionalnog grupiranja na mnogim različitim skupovima podataka. Postoji nedostatak SOM-e koja se može izvoditi na unaprijed definiranom dovršenom skupu podataka. Na vremenskim tokovima skupova podataka pojavljuju se razni problemi prilikom grupiranja pomoću standardne SOM-e jer se vremenski tokovi podataka generiraju ovisno o vremenu. U ovoj studiji značajka kliznog prozora uključena je u standardnu SOM-u za grupiranje vremenskih tokova podataka. Stoga, kombinacija SOM i kliznog prozora (SOM + SW) daje točnije rezultate prilikom grupiranja podataka na vremenskom toku skupova podataka. Da bi se to dokazalo, testiran je skup podataka o uporabi interneta mobilnog operatora u Turskoj. Uzeti skup podataka mobilnog operatera grupiran je prema klasičnoj SOM-i, a zatim je procijenjena buduća uporaba podataka pretplatnika. Isti skup podataka primijenjen je na SOM + SW za izvođenje istih simulacija.

Ključne riječi: *grupiranje; klizni prozor; mobilni operateri; samoorganizirane mape (SOM); vremenski tok skupova podataka*

1 Introduction

In the last decade, academic or industrial information has been rising at exceptional rates. Parsing new information from gigantic databases is challenging, expensive and time consuming if done routinely. The key objective is to find consistencies and relations in the data, thus gaining access to hidden and potentially suitable data. The Self-Organizing Map (SOM) is a properly famous neural network and certainly one of the most popular unsupervised learning algorithms. Since its invention by Finnish Professor Teuvo Kohonen in the early 1980s, more than 4000 research articles have been published on the algorithm, its conception and uses [1, 2]. The SOM mapping is preserving, namely the more similar two data samples are in the input space, the closer they will appear together on the final displayed map. This allows the user to identify clusters such as large sets of a specific type of input pattern.

There are many studies improving SOM algorithm to solve a specific problem. In one of these studies, Chaudhary et al. (2014) modified the classical SOM in a way that as well as the farthest and nearest neurons from the winner neuron, the winning frequency of each neuron was taken into account for updating the weight [3]. In another study, Ghaseminezhad and Karami (2011) presented a novel SOM-based algorithm for clustering discrete groups of data and they indicated the classic SOM algorithm could not cluster discrete data correctly [4]. In some studies, they used SOM algorithm combined with other methods such as recurrent prediction [5] for times series, genetic algorithm [6] for data visualization, Markov Model [7] for biological sequence analysis, and support vector machine [8] for classification of enzymes controlling cell division. In addition, Sliding Window has many usages for Neural Network. In one of these studies Steven F. B. (1979) uses sliding window to calculate windowed speech data for suppression of acoustic noise. In this article the weight of input data in sliding window is used to calculate weight of clusters dynamically [9]. In fact, Neural Networks have been widely used as time series forecasters. In one of these studies, Frank, R. J., Neil, D., and Stephen, P. H. (2001) attempted to answer the question "can the performance of sliding window feed-forward neural network predictors be optimized using theoretically motivated heuristics". They use ATM network traffic data. They calculate the relationship between datasets and network performance [10]. In this study, Sliding Window feature is used for hourly period and calculated neighborhood of clusters. The SOM is considered with Sliding Window that is called 'Sliding Window' (SOM+SW) approach that provides dynamical generated time-stream data clusters.

In this paper, Section 2 of the paper introduces the SOM basics and its working. In Section 3, the possessed different features by using SOM+SW technique are discussed. Section 4 gives an evaluation of the system briefly. Section 5 discusses a case study for application of the SOM and SOM+SW to a time-based Dynamic Quota Calculation System (DQCS). Section 6 presents the retrieved simulation results for different constant ranges while Section 7 is dedicated to the conclusion.

2 Classical Self-Organizing Maps (SOM)

The basic self-organizing system is a one- or two dimensional array of neurons in the form of neighboring units. The first simulation study related to SOM ordering

process was performed by Kohonen [11]. SOM is used for many practical applications as a clustering method [12, 15]. The basic algorithm of SOM neural network is as follows [13, 14]:

1) *Each node's weights are initialized randomly.*
2) *A vector is chosen at random from training data and attended to the lattice.*
3) *Every node is tested to calculate which one weights closer to input vector. Best Matching Unit (BMU) is the successive node.*
4) *The radius of the neighborhood of the BMU is calculated. Nodes within the range of radius are defined as to be inside the BMU's neighborhood.*
5) *Each neighboring node's (the nodes found in step 4) weights are adjusted to make them more like the input vector. The closer a node is to the BMU, the more its weights get altered.*
6) *Repeat step 2 for N iterations.*

A practical SOM is a two-layer feed-forward Artificial Neural Network (ANN). The input layer consists of the neurons indicating the attributes used for clustering. The output layer stands for the clusters usually arranged in the form of hexagonal or rectangular grid [13]. There is a reference vector for each cluster neuron in order to indicate the weights between input neurons and the related cluster neuron. SOM algorithm consists of two parts: Training and Mapping. In training part, an unsupervised learning algorithm combined with a neighborhood function is used to determine the reference vectors. Finally, the input rows are applied to SOM to construct the cluster map. In this study, the reference vector is weight of the cluster. After preparation, dataset is entered to the system and changed the average weights of the clusters. The steps depict the working mechanism of SOM:

1) *Start SOM.*
2) *Clusters are prepared with starting conditions. For these conditions the weight of the clusters is given randomly. The counts of the clusters must be predefined also. In this study, the maximum number of the clusters is 100 (this information is real data that is taken from one of the lead mobile companies in Turkey[1]).*
3) *The training data is taken into account.*
4) *The closest cluster is defined.*
5) *The input is added in that defined cluster and average weight is changed.*
6) *This process is repeated continuously until clusters have no significant change.*
7) *If no significant change, the clusters are ready for the real dataset.*
8) *The dataset is taken into account.*
9) *The closest cluster is defined.*
10) *The input is added in that defined cluster and average weight is changed.*
11) *This process is repeated continuously until clusters have no significant change.*

In Step 7, the used data set is not a time stream data set since SOM is not suitable on the time-stream data set. The results and added dynamism feature to the SOM are discussed in detail in the next sections.

3 The proposed Self-Organizing Maps with Sliding Window (SOM+SW)

As it is mentioned, the disadvantage of the SOM is that the system will lose its dynamism in time, namely, the weight of cluster is not affected when one more new weighted data is added as an input to the cluster if former inputs are stored in the same cluster forever).

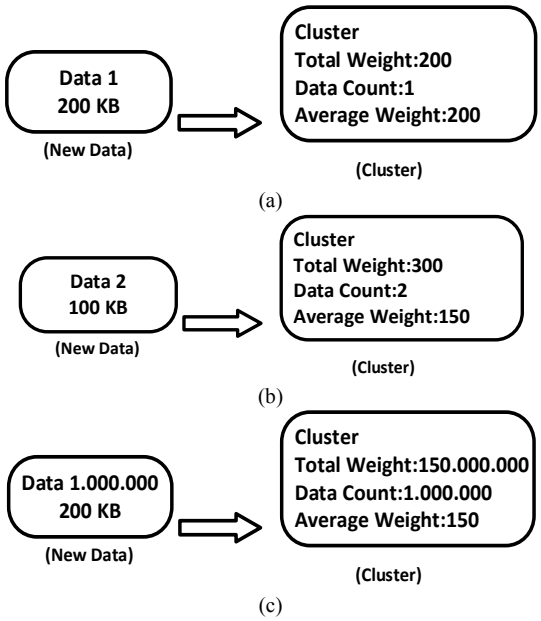

Figure 1 Data flow for SOM Clustering: (a) First data into classical SOM cluster (b) Second data into classical SOM cluster (c) After 10^6th data into classical SOM cluster.

An example: as shown in Fig. 1(a) to 1(c), continuous long term arrivals of incoming data packages will lead to lose dynamism (cluster weighting average value) of clusters according to the classic SOM clusters. Let the average weight of a cluster shown in Fig. 1(a) be approximately 200 KB. In Fig. 1(b), the 2nd data of 100 KB is included to the same cluster that has affected the cluster in 50% weight. But as it is seen in Figure 1(c), there is not any effect on average weight of the same cluster when 10^6th incoming data with 200 KB weight is included to the same cluster. The cluster still has the same average weight that is 150 KB and loses its dynamism since the cluster's average weight value will always stay fixed. In order to solve this problem, Sliding Window sense has been added into classical SOM neural network (SOM+SW).

According to the extended SOM with Sliding Window logic, the last incoming data is included into the cluster in this situation as well. However, it is excluded from that cluster after a specific time zone (for instance, after 1 hour). A flow diagram of the proposed SOM+SW mechanism is presented below in Fig. 2. After a new input is entered to the system the flow is started. The closeness weight is defined regarding the average weight values of clusters created by SOM+SW mechanism. Therefore, the cluster with the closest average weight value is founded. If the calculated distance weight to the closest cluster's

[1] http://www.avea.com.tr

average weight is lower than a range value, then the new data is added into that cluster. In addition, the average weight value of that cluster is updated. If the distance to the closest cluster's average weight value is higher than the range value, then a new cluster is created by using that usage data. After 1 hour, weight of this input is excluded from that cluster and also average weight of the cluster is updated.

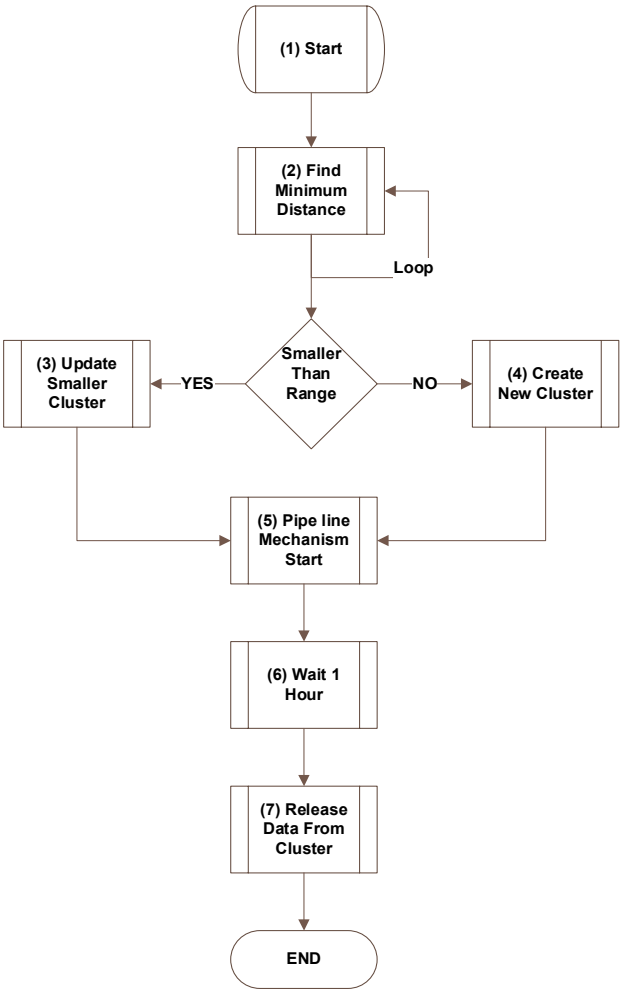

Figure 2 SOM+SW mechanism.

a) The weight of incoming input is taken into account (depicted in Step 1 of Fig. 2).
b) The weight of closest cluster for the input is found (depicted in Step 2 of Fig. 2).
c) If the average weight of the found closest cluster is closer than a predefined threshold value (assumed as a constant range i.e. 500) then the input is included to the cluster (depicted in Step 3 of Fig. 2).
d) If the average weight value of the found closest cluster is further than the threshold value then a new cluster is created. Therefore, the number of clusters is acquired dynamically in this way (depicted in Step 4 of Fig. 2).
e) After all, Sliding Window approach is used in order to gain dynamism to the average weights of each generated clusters. To do this, an input which was included in a cluster over an hour is needed to remove from that cluster. In this way, generated clusters stay always dynamic since the average weights of these clusters are changed according to inputs weight within 1 hour time slots (depicted in Step 5, 6 and 7 of Fig. 2).

Next section depicts obtained better results through SOM + SW as shown in the above clustering algorithm that is performed on different time-stream data sets as for the internet usages of subscribers of a Mobile operator case study.

4 A case study for Mobile Data Communication Systems (MDCS) through SOM and SOM+SW

In this section, the classical SOM and SOM+SW approaches are compared by simulating the dynamic quota allocations and charging problem for internet users in Mobile Data Communication Systems (MDCS). Therefore, a case study is initiated and various simulations performed on a real data set about internet usages of mobile subscribers. The real data set belongs to one of the leading mobile companies[1] in Turkey. In this case study age, gender, home city, client profile (CRM segment), tariff values of subscribers, and their instant internet usages (in terms of weight, KB) are considered as parameters. Nowadays, in MDCS, a constant quota size is assigned for internet usages to mobile customers without regard whether the subscriber has high or low data usage. In general, the 750 KB quota size is assigned to subscribers by the charging system of MDCS statically. The arrangement of instant dynamic instant quota size only with respect to the subscribers with low data usage causes various performance problems such as heavy control signalization. On the other side, the arrangement of quota size only with respect to the subscribers with high data usage leads to unnecessary quota allocation. Because of these reasons, a dynamic quota allocation method is required to increase performance of MDCSs. Therefore, SOM and SOM+SW approaches are simulated to estimate the future total data usage of the subscribers to perform dynamic quota allocation.

Firstly, classical SOM is evaluated in terms of the amount of past internet usage, age, gender, home city, client profile (CRM segment) and tariff. Then, the mechanism of classical SOM is combined with Sliding Window logic to perform the SOM+SW. After clustering, the subscriber characteristics used to estimate their future data usages are: Age, Tariff, Gender, City, and CRM_SEGMENT (Tab. 1).

- Call Start Date: When a user starts to use mobile data,
- Total Data Usage: Total data usage of a subscriber after completing his/her internet usage,
- Birth Date: Birthdate of a subscriber,
- Gender: Gender of a subscriber,
- City: Location of a subscriber,
- Current CRM Segment: Classification of a subscriber that is assigned by the company,
- Tariff: Bought tariff type by a subscriber.

A 3-step general procedure is followed when the SOM and SOM+SW approaches are compared in the case study:

- Calculate amount of Estimated Data Usage (will be discussed in below for SOM and SOM+SW respectively).

- Get amount of the Real Data Usage (Total Data Usage) from the dataset (Tab. 1).
- Calculate the Difference of Accuracy on Data Usage= |amount of Real Data Usage - amount of Estimated Data Usage|.

Table 1 A portion of internet data usages between 01/06/2012 and 30/06/2012 for a MDCS is presented during simulations.

Call Start Date	Total Data Usage	Birth Date	Gender	City	Current CRM Segment	Tariff
10.06.2012 21:31	4869	28.10.1960	M	GIRESUN	Bronze	Plan Voucher
24.06.2012 13:05	156	28.10.1960	F	ISTANBUL	Gold	Plan Voucher
13.06.2012 04:57	3906	13.03.1969	M	ANKARA	Bronze	Free Voice Call
10.06.2012 21:31	4869	28.10.1960	M	GIRESUN	Bronze	Plan Voucher
24.06.2012 13:05	156	28.10.1960	F	ISTANBUL	Gold	Plan Voucher
13.06.2012 04:57	3906	13.03.1969	M	ANKARA	Bronze	Free Voice Call
28.06.2012 23:57	11723	25.09.1990	M	ADANA	Bronze	Free Video Call
28.06.2012 23:59	3028	25.09.1990	F	SAMSUN	Silver	Student
29.06.2012 00:00	9008	12.08.1980	F	ORDU	Silver	Plan Voucher
29.06.2012 00:05	364	11.09.1985	F	IZMIR	Bronze	Plan Voucher
29.06.2012 00:09	4950	11.09.1985	M	KONYA	Silver	Plan Voucher
29.06.2012 00:14	1070	12.08.1980	F	ISTANBUL	Gold	Free Video Call
29.06.2012 06:51	3763	06.10.1975	F	ORDU	Bronze	Plan Voucher
06.06.2012 16:33	44	10.03.1972	M	ISTANBUL	Bronze	Student
06.06.2012 16:37	130	20.10.1962	M	ISTANBUL	Bronze	Plan Voucher

Procedure A. Simulation of SOM provides the information about the difference between the Real Data Usages (Total_Data_Usage) and the Estimated Data Usage (namely, Difference for SOM):

- Apply SOM to generated clusters according to Total Data Usage of the dataset (Table 1). Here, the clusters are generated according to Real Data Usages (Total Data Usage) of entire users in dataset (Table 1).
- Each user in the data set is evaluated according to similarity in terms of Age, City, CRM Segment, and Tariff parameters to find a cluster which has the max number of the similar users for that user.
- Then, the average weight of the found cluster is assigned as an Estimated Data Usage of each user in dataset (Table 1). Amount of Estimated Data Usage of each user is defined here.
- Finally, absolute Difference of Accuracy on Data Usage (Difference for SOM) for each user in dataset is calculated by subtracting the amount of Real Data Usages (Total_Data_Usage) from amount of Estimated Data Usage of that user.

Procedure B. Simulation of SOM + SW provides the information about the difference between the Real Data Usages (Total_Data_Usage) and the Estimated Data Usage (namely, Difference for SOM+ SW):

- Apply SOM+SW to generate the first cluster for the first arrived user to the system (in Tab. 1). Here, the first cluster is generated according to Real Data Usages (Total_Data_Usage) of that first user (in Table 1). Then, the system generates m-clusters after arriving of n-users up to now. According to the SOM+SW algorithm in Fig. 2, the generated clusters kept the users who had arrived to the system in the last one hour. The clusters are ready to Estimated Data Usage of a new incoming user to the system at the moment.
- When a new user has arrived to the system, the system tries to find a cluster according to similarity in terms of Age, City, CRM Segment, and Tariff parameters which has the max number of similar users for the new arrived user.
- Then, the average weight of the found cluster is assigned as an Estimated Data Usage of the new arrived user in dataset (Tab. 1). Amount of Estimated Data Usage of the new arrived user is defined here.
- Finally, absolute of the Difference of Accuracy on Data Usage (Difference for SOM+SW) for the new arrived user in dataset is calculated by subtracting the amount of Real Data Usages (Total_Data_Usage) from amount of Estimated Data Usage of that user.
- The Real Data Usages (Total_Data_Usage) of the new arrived user in dataset (Tab. 1) is used for updating clusters (the aim of this is to keep new arrived users in clusters who have arrived in last one hour) after completing the calculating of the Difference of Accuracy on Data Usage (Difference for SOM+SW).

The retrieved results of SOM+SW are found better than SOM results on the same datasets for the same problem above that are depicted thru result graphs in the next section

5 Evaluations

The complete data set about data usages of subscribers for the Mobile company1 between 01/06/2012 and 30/06/2012 is considered in simulations (a small portion of the entire dataset is depicted in Table 1). The complete dataset is separated into four different equal datasets which are listed below;

- Dataset 1 involves the data between 01/06/2012 00:00AM and 07/06/2012 23:59PM (contains 5436 data)
- Dataset 2 involves the data between 08/06/2012 00:00AM and 14/06/2012 23:59PM (contains 4027 data)
- Dataset 3 involves the data between 15/06/2012 00:00AM and 21/06/2012 23:59PM (contains 4984 data)
- Dataset 4 involves the data between 22/06/2012 00:00AM and 30/06/2012 23:59PM (contains 6450 data)

The estimated results of "SOM (Procedure A)" and "SOM+SW (Procedure B)" for the given above four

different datasets are presented in the second and third columns in Tab. 2 (for Dataset 1), Tab. 3 (for Dataset 2), Tab. 4 (for Dataset 3) and Tab. 5 (for Dataset 4) respectively. The calculated data are assembled based on 6 hour periods in the "Time" column of these tables.

The first column depicts the date format that is YYYYMMDDHH (i.e. 2012060100 means 01/06/2012: 00 AM). In the second column, the sum of differences between Total Data Usage and Estimated Data Usage of each user according to the SOM is presented. In the third column, the sum of differences between Total Data Usage and Estimated Data Usage of each user according to the SOM+SW is presented.

At the bottom of these tables, the sum of the differences is presented. It can be seen that the difference between the Total Data Usage and the Estimated Data Usage by using the SOM+SW is found lower than the difference of the SOM.

Table 2 This table depicts the retrieved results of "Difference for SOM (Procedure A)" and "Difference for SOM+SW (Procedure B)" of the first week (Dataset 1). The data was grouped for 6 hours periods in Dataset 1 that is highlighted bold in the "Time" column.

Time	Difference for SOM (Procedure A)	Difference for SOM+SW (Procedure B)
2012060100	213183002	136559380
2012060106	141084661	51818300
2012060112	308611313	210833876
2012060118	362551688	184404176
2012060200	153996011	57836418
2012060206	153993937	17573359
2012060212	154131584	153750864
2012060218	158003349	51533445
2012060300	110272033	48857874
2012060306	244397233	94196029
2012060312	282329558	216386145
2012060318	433706408	310018882
2012060400	229368314	184718527
2012060406	191300232	170276288
2012060412	182979067	168940899
2012060418	256552705	173092733
2012060500	145694301	132377366
2012060506	152091065	57542835
2012060512	185844812	184173092
2012060518	303867838	217456826
2012060600	214455694	47557837
2012060606	257588022	64659354
2012060612	198806002	219398139
2012060618	241617620	159152071
2012060700	157338770	29361494
2012060706	90596732	35939463
2012060712	92311914	37324454
2012060718	266355883	272315338
TOTAL	5.883.029.748	3.688.055.464

5.1 Simulation results on Dataset 1

The accuracy is increased by 38% for Dataset 1 by considering the SOM+SW approach while estimating future data usage of subscribers. The 38% is calculated from the formula ((1 − Differences for SOM+SW /Differences for SOM)×100) that is ((1 − 3688055464/5883029748)×100) in this case. Then, Fig. 3 is obtained from Dataset 1. As shown in the graph, the SOM+SW algorithm gives more lucrative results than the results of the SOM. While the X axis in this graph refers to "Time" of Tab. 2, the Y axis represents "Difference for SOM (Procedure A)" and "Difference for SOM+SW (Procedure B)" results of Tab. 2.

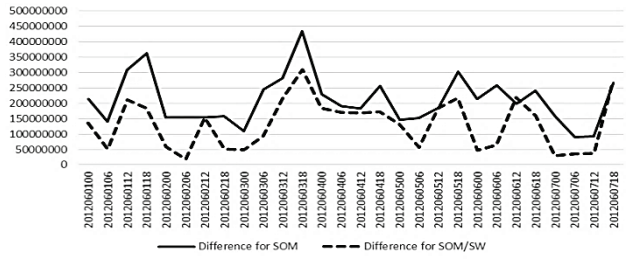

Figure 3 Graph depicts the comparison of the SOM and SOM+SW for entire Dataset 1.

5.2 Simulation results on Dataset 2

Table 3 This table depicts the retrieved results of "Difference for SOM (Procedure A)" and "Difference for SOM+SW (Procedure B)" of the second week (Dataset 2).

Time	Difference for SOM (Procedure A)	Difference for SOM+SW (Procedure B)
2012060800	119784210	80828513
2012060806	115188489	11531542
2012060812	175446253	79905370
2012060818	218929629	119640166
2012060900	111363105	62949485
2012060906	192176845	205110127
2012060912	268802411	206552736
2012060918	209081779	210873391
2012061000	161805410	29073418
2012061006	169023277	73905783
2012061012	182927461	36472832
2012061018	160688386	29006525
2012061100	173394329	33343268
2012061106	138848047	15040155
2012061112	169465429	53242709
2012061118	234435503	91945180
2012061200	172831827	26429254
2012061206	119327937	19459952
2012061212	105997720	19419166
2012061218	184225517	39537524
2012061300	131532465	52636251
2012061306	105161380	17873286
2012061312	158908533	43533067
2012061318	249836429	125789664
2012061400	148116330	178838837
2012061406	115075686	21764241
2012061412	105771310	21933410
2012061418	162247789	63177960
TOTAL	4.560.393.486	1.969.813.812

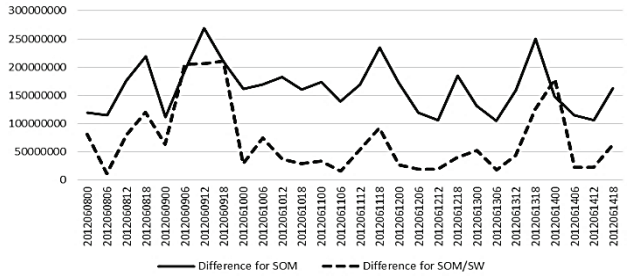

Figure 4 Graph depicts the comparison of the SOM and SOM+SW for Dataset 2.

It can be seen that the accuracy is increased by 57% for Dataset 2. The following graph in Fig. 4 is obtained from Dataset 2.

5.3 Simulation results on Dataset 3

It can be seen that the accuracy is increased by 3% for Dataset 3 by considering the SOM+SW approach while

estimating future data usage of subscribers. The following graph in Fig. 5 is obtained from the Dataset 3.

Table 4 This table depicts the retrieved results of "Difference for SOM (Procedure A)" and "Difference for SOM+SW (Procedure B)" of the third week (Dataset 3).

Time	Difference for SOM (Procedure A)	Difference for SOM+SW (Procedure B)
2012061500	129113091	122711746
2012061506	192832868	59324036
2012061512	201766511	244448829
2012061518	379627716	284467989
2012061600	215330449	181969707
2012061606	135413561	16232375
2012061612	276735195	180284326
2012061618	386825353	519088293
2012061700	226611996	193301518
2012061706	138646855	21738451
2012061712	286816829	173857404
2012061718	296343173	273198710
2012061800	189716427	189447546
2012061806	175132966	119041705
2012061812	163724401	81561621
2012061818	279251875	431254324
2012061900	157243244	71044431
2012061906	205814198	45568311
2012061912	288315923	315117670
2012061918	366632832	742713424
2012062000	251260088	159275013
2012062006	216018816	40307919
2012062012	339557870	468812084
2012062018	331692915	334198136
2012062100	213845894	414556670
2012062106	209350335	42897431
2012062112	289669552	317906517
2012062118	385621236	732311908
TOTAL	6.928.912.169	6.776.638.094

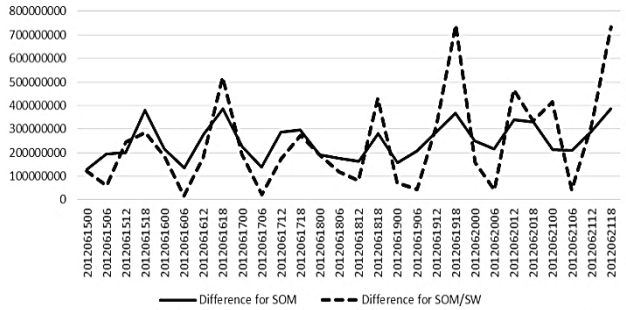

Figure 5 Graph depicts comparison of the SOM and SOM+SW for Dataset 3.

5.4 Simulation results on Dataset 4

It can be seen that the accuracy is increased by 17% for Dataset 4 by considering the SOM+SW approach while estimating future data usage of subscribers. The following graph in Fig. 6 is obtained from the Dataset 4.

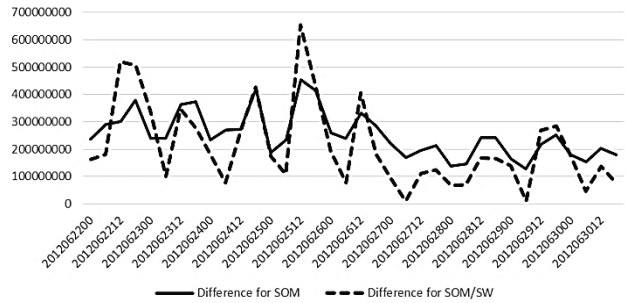

Figure 6 Graph depicts comparison of the SOM and SOM+SW for Dataset 4.

Table 5 This table depicts the retrieved results of "Difference for SOM (Procedure A)" and "Difference for SOM+SW (Procedure B)" of the fourth week (Dataset 4).

Time	Difference for SOM (Procedure A)	Difference for SOM+SW (Procedure B)
2012062200	235553931	162044291
2012062206	292171996	180341622
2012062212	302464338	519477809
2012062218	379887232	507856991
2012062300	239349240	339398485
2012062306	239724743	102158903
2012062312	362521330	345422415
2012062318	373875244	276824190
2012062400	234664989	177849618
2012062406	271354981	78221492
2012062412	273181410	267907949
2012062418	420574802	426450536
2012062500	188198222	173775888
2012062506	233977994	109847994
2012062512	453113897	654465865
2012062518	414504887	426333104
2012062600	259575442	193424842
2012062606	239163297	77673509
2012062612	332160232	406219865
2012062618	285338043	181304603
2012062700	221220768	94145838
2012062706	168238430	10205081
2012062712	196682109	112317358
2012062718	213947067	123904811
2012062800	137199305	66421807
2012062806	146125471	68653527
2012062812	240848636	168396671
2012062818	242421627	165285296
2012062900	164810059	139279987
2012062906	129024949	10332312
2012062912	218116338	269965016
2012062918	252831176	283910688
2012063000	179549309	173459243
2012063006	154043049	47476099
2012063012	204387677	137140083
2012063018	181163279	77407117
TOTAL	9.081.965.499	7.555.300.905

5.5 Final evaluation for simulation results of four datasets

Table 6 The table presents the results of four different datasets.

Datasets	Difference for SOM*	Difference for SOM+SW*	Correctness*
Dataset 1: between 01/06/2012 and 07/06/2012	5.883.029.748	3.688.055.464	38%
Dataset 2: between 08/06/2012 and 14/06/2012	4.560.393.486	1.969.813.812	57%
Dataset 3: between 15/06/2012 and 21/06/2012	6.928.912.169	6.776.638.094	3%
Dataset 4: between 22/06/2012 and 30/06/2012	9.081.965.499	7.555.300.905	17%
TOTAL	26.454.300.902	19.989.808.275	24.5%

* The fractional parts of the given values above are ignored.

As a result of simulations, entire retrieved results of the "Difference for SOM (Procedure A)" and "Difference for SOM+SW (Procedure B)" are depicted in Tab. 6. The retrieved results are: 26.454.300.902 bytes for the "Difference for SOM" and also 19.989.808.275 bytes for "Difference for SOM+SW". The results depict that the

accuracy of the SOM+SW is by 24.5% better than the SOM result.

6 Evaluation according to different constant ranges

The number of clusters is acquired dynamically in SOM+SW. In addition, different size of threshold values is considered during the simulations to understand the effects of the constant size on the number of clusters. The other considered threshold values except 500 kB are: 125 kB, 250 kB, 1000 kB and 2000 kB. The threshold values are applied on each above four datasets that are listed below:

- *Difference of SOM = |amount of Real Data Usage - amount of Estimated Data Usage for whole Datasets.*
- *Difference of SOM+SW (125) = amount of Real Data Usage - amount of Estimated Data Usage with 125 kB constant range for whole Datasets.*
- *Difference of SOM+SW (250) = amount of Real Data Usage - amount of Estimated Data Usage with 250 kB constant range for whole Datasets*
- *Difference of SOM+SW (1000) = amount of Real Data Usage - amount of Estimated Data Usage with 1000 kB constant range for whole Datasets*
- *Difference of SOM+SW (2000) = amount of Real Data Usage - amount of Estimated Data Usage with 2000 kB constant range for whole Datasets.*

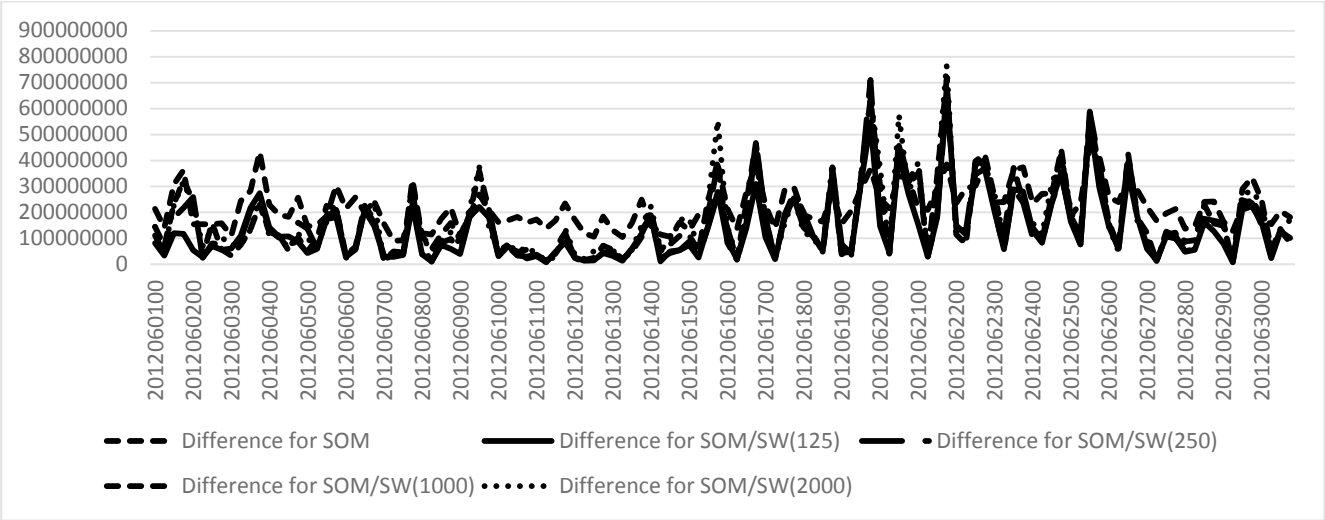

Figure 7 Graph depicts the comparison of the SOM and SOM+SW (with 4 different threshold values; 125 kB, 250 kB, 1000 kB, and 2000 kB)

Fig. 7 depicts the retrieved results after applying the different threshold values on whole Datasets. While the *X* axis in Fig. 7 refers to "Time", the *Y* axis represents "Difference for SOM (Procedure A)" and also calculated the "Difference for SOM+SW (Procedure B)" with these ranges: 125 kB, 250 kB, 1000 kB and 2000 kB. In Fig. 7, the SOM+SW results give more lucrative results than the result of SOM for Datasets.

Table 7 The table presents simulation results of whole datasets with five different constant values (125 kB, 250 kB, 500 kB, 1000 kB, and 2000 kB).

Dataset	SOM*	SOM+SW with 125 range*	SOM+SW250 with range*	SOM+SW with 500 range*	SOM+SW with 1000 range*	SOM+SW with 2000 range*
Whole Datasets	26.454.300.902	15.897.928.338	18.230.114.112	19.989.808.275	19.951.421.662	20.486.376.521
CORRECTNESS		40%	31%	24%	24%	22%

* The fractional parts of the given values above are discarded.

- When threshold value is 125 kB, 40% Better correctness for Datasets (The 50% is calculated from the formula ((1 − Differences for SOM+SW /Differences for SOM)×100)
- When threshold value is 250 kB, 31% Better correctness for Datasets.
- When threshold value is 500 kB, 24% Better correctness for Datasets.
- When threshold value is 1000 kB, 24% Better correctness for Datasets.
- When threshold value is 2000 kB, 24% Better correctness for Datasets.

6.1 Observations and comparative studies

As seen in Tab. 7 different threshold values can be used for calculations. These values can be varied. The smaller threshold gives better results than bigger thresholds.

There are several other algorithms which are similar to SOM. For example, *k*-means is a clustering algorithm which aims to cluster n data into *k* clusters in which each data belongs to the cluster with the nearest mean [16]. In order to dynamically change the number of clusters, *X*-means clustering algorithm has been developed over *k*-means [17]. It is possible to use sliding window in *X*-means algorithm as in SOM in a way that any data at any cluster can be removed at the end of windows time period and the related cluster weight can be updated dynamically. Unlike SOM, in order to create a new cluster, an old cluster must be divided into two parts in *X*-means. In this case, two new created clusters are close to each other. This can be a disadvantage for *X*-means because new data may have little relation with this cluster when very different data not belonging to any cluster

occurs. In SOM+SW, diameter is used for creating new clusters. When a new data which does not belong to any cluster comes to system and its distance is bigger than the closest cluster, SOM+SW creates a new cluster.

7 Conclusion

In this article, an extended SOM algorithm with the Sliding Window (SOM + SW) approach is proposed and compared with classical SOM via various performed simulations based on a real data set about internet usages of mobile subscribers. The data set is taken from one of the lead mobile companies in Turkey[1]. In this study, the Sliding Window feature is added to the classical SOM by recalculating the average weight of each cluster for a specific time period. In order to figure out that SOM+SW gives more accurate results for clustering on time-stream data sets, a set of internet usage data from the mobile operator in Turkey is used as a case study. By using the past data usages of subscribers in this dataset, the clusters where the subscribers are involved have been determined for SOM and SOM+SW. After clustering, the subscriber characteristics Age, Tariff, Gender, City, and CRM_SEGMENT are used to estimate their future data usages. However, during the SOM+SW simulations, only last one hour data of the data set is used in generated clusters because of Sliding Window feature. In addition, SOM+SW is simulated for different threshold value parameter such as 125 kB, 250 kB, 500 kB, 1000 kB and 2000 kB. As a conclusion, the SOM+SW always outperforms SOM in terms of the difference between real and estimated data usage for all range values by giving more accuracy for small values due to the better cluster assignments for subscribers.

Acknowledgements

The article is presented from a funded project that is the cooperation with AVEA Mobile company[1] and Istanbul Aydin University[2] with the support of the Ministry of Science, Industry and Technology[3] (SANTEZ funded program- project number: 00874.STZ.2011-1) in Turkey.

8 References

[1] Kohonen, T. The self-organizing Map. // Proceedings of the IEEE, 78, 9(1990), pp. 1464-1480. https://doi.org/10.1109/5.58325

[2] Kohonen, T. Self-Organized Formation of Topologically Correct Feature Maps. // Springer-Verlag Biological Cybernetics. 43, 1(1982), pp. 59-69. https://doi.org/10.1007/BF00337288

[3] Chaudhary, V.; Bhatia, R.S.; Ahlawat, A. K. A Novel Self-Organizing Map (SOM) Learning Algorithm with Nearest and Farthest neurons. // Alexandria Engineering Journal. 53(2014), pp. 827-831. https://doi.org/10.1016/j.aej.2014.09.007

[4] Ghaseminezhad, M. H.; Karami, A. A novel Self-organizing Map (SOM) Neural Network for Discrete Groups of Data Clustering. // Applied Soft Computing. 11(2011), pp. 3771-3778. https://doi.org/10.1016/j.asoc.2011.02.009

[5] Cherif, A.; Hubert, C.; Romuald, B. SOM Time Series Clustering And Prediction with Recurrent Neural Networks. // Neurocomputing. 74, 11(2011), pp. 1936-1944. https://doi.org/10.1016/j.neucom.2010.11.026

[6] Sirisin, S.; Jonburom, W.; Rattanakorn, N.; Pornsuwancharoen, N. A New Technique Gray Scale Display of Input Data Using Shooting SOM and Genetic Algorithm. // Procedia Engineering. 32(2012), pp. 556-563. https://doi.org/10.1016/j.proeng.2012.01.1308

[7] Ferles, C.; Andreas, S. Self-Organizing Hidden Markov Model Map (SOHMMM). // Neural Networks. 48(2013), pp. 133-147. https://doi.org/10.1016/j.neunet.2013.07.011

[8] Yan, A.; Nie, X.; Wang, K.; Wang, M. Classification of Aurora Kinase Inhibitors by Self-Organizing Map (SOM) and Support Vector Machine (SVM). // European Journal of Medicinal Chemistry. 61(2013), pp. 73-83. https://doi.org/10.1016/j.ejmech.2012.06.037

[9] Steven, F. B. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. // Acoustics, Speech and Signal Processing, IEEE Transactions on. 27, 2(1979), pp. 113-120. https://doi.org/10.1109/TASSP.1979.1163209

[10] Frank, R. J.; Neil, D.; Stephen, P. H. Time Series Prediction and Neural Networks. // Journal of Intelligent and Robotic Systems. 31, 1-3(2001), pp. 91-103. https://doi.org/10.1023/A:1017469712308

[11] Kohonen, T. Self-organized Formation of Topologically Correct Feature Maps. // Biological Cybernetics. 43, 1(1982), pp. 59-69. https://doi.org/10.1007/BF00337288

[12] Tulankar, K.; Manali, K.; Rakhi, W. Clustering Telecom Customers using Emergent Self Organizing Maps for Business Profitability. // International Journal of Computer Science and Technology. 3, 1(2012), pp. 256-259.

[13] Kohonen, T. Self-Organizing Maps, 3rd Edition. Springer-Verlag, Berlin, 2001. https://doi.org/10.1007/978-3-642-56927-2

[14] Fausett, L. Fundamentals of Neural Networks. Prentice Hall, New Jersey, 1994.

[15] Seret, A.; Verbraken, T.; Versailles, S.; Baesens, B. A New SOM-based Method for Profile Generation: Theory and an Application in Direct Marketing. // European Journal of Operational Research. 220, 1(2012), pp. 199-209. https://doi.org/10.1016/j.ejor.2012.01.044

[16] MacQueen, J. B. Some Methods for Classification and Analysis of Multivariate Observations. // Proceedings of the Fifth Symposium on Math, Statistics, and Probability / Berkeley, University of California Press, 1967, pp. 281-297.

[17] Pelleg, D.; Moore, A. W. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. // ICML'00 Proceedings of the Seventeenth International Conference on Machine Learning / Stanford University, 2000, pp. 727-734.

Authors' addresses

Ulaş Çelenk, PhD.
Istanbul University,
INNOVA, ITU Ayazaga Campus Teknokent ARI-4,
Maslak, Istanbul, Turkey
E-mail: ulascelenk@gmail.com
E-mail: ucelenk@innova.com.tr

Duygu Çelik Ertuğrul, Assoc. Prof. Dr.
Eastern Mediterranean University,
Engineering Faculty,
Computer Engineering Department,
Famagusta, North Cyprus, via Mersin -10, Turkey

[2] http://www.aydin.edu.tr/index_eng.asp

[3] http://www.sanayi.gov.tr/Default.aspx?lng=en

E-mail: duygu.celik@emu.edu.tr
E-mail: duygucelik@msn.com

Metin Zontul, Asst. Prof. Dr.
Istanbul Aydin University,
Faculty of Engineering,
Software Engineering Dept.
Halit Aydın Campus No: 38, Sefaköy–Küçükçekmece, Istanbul,
34295, Turkey
E-mail: metinzontul@aydin.edu.tr

Osman Nuri Uçan, Prof. Dr.
Istanbul Aydin University,
Faculty of Engineering,
Electrical & Electronics Engineering Dept.,
Halit Aydın Campus No: 38, Sefaköy–Küçükçekmece, Istanbul,
34295, Turkey
E-mail: uosman@aydin.edu.tr