International Journal of Computing Science and Information Technology, 2018
ISSN: 2278-9669,April 2018 (http://ijcsit.org)

eARMIA model to Analyses and Ranking Attributes of Time Series Data

S.Gokila[1], Dr. K. Ananda Kumar[2], Dr. A. Bharathi[3]

[1] Research Scholar, Bharathiar University, Coimbatore, gokilaphd@gmal.com
[2] Department of Computer Science and Engineering, Bannariamman Institute of Technology, Erode, India, anandakumark@bitsathy.ac.in
[3] Department of IT, Bannariamman Institute of Technology,Erode,India, bharathia@bitsathy.ac.in

Abstract

The objective of the proposed algorithm is to select the attribute's rank index based on the involvements of the same in data set; identifying the global and local outline of entire data set. Because the attributes in data set, number of cluster and outline object are major intervention decide the accuracy of each clusters. The attribute ranking of proposed method automate the number of clusters and the center of the cluster. The ranking over caps the attributes with necessary weighted one. This transformation of indexed attribute decide the cluster accuracy. The uniqueness of the algorithm is exhibited in maintaining complete set of attributes without eliminating the low ranked attribute which are required for predictive analysis and in grouping the outline data set as clusters.

Key Words : ARIMA, Indexing, Time Series, Ranking

1. Introduction

The dataset of some domain consists more sensitive set of values. That sensitivity extraction is possible when dataset are sliced occurring to the relevance. Another key note to be considered is the number of attributes involved in unsupervised mining. Automated selections of number of cluster are another challenge. All these are handled in proposed model. The attributes are selected based on the eARIMA analysis result; number of clusters are equal to the number of attributes in the data set. These method of decision helps to get the pattern based on the particular attribute. The variation in the basic projected space clustering is important in data set like time series data. Because the dimension removed as outline may influence a predication of subsequent values [2,5,7]. For example during the early time series may have less reading but that is also one of the value decides the value of next upcoming time data set in prediction. So the attributes must be kept as such even it is less important in that cluster. The model analyzed and transformed using eARIMA are applied with MPSKM to ensemble the data with all level of attributes [8].

1.1. ARIMA Model

Generalized random walk models fine-tuned to eliminate all residual autocorrelation. The model takes exponential smoothing models that can corporate long-term trends and seasonality. It is a stationarized regression models that use lags of the dependent variables and/or lags of the forecast errors as regressors. The most general class of forecasting models for time series that can be stationarized by transformations such as differencing, logging, and or deflating. The ARIMA called an "ARIMA(p,d,q)" model include one constant. The parameters of these model are number of auto regression (p), number of non seasonal regression (d), number of moving average (q).

In Arima Model, the general forecasting equation is

$$\hat{y}_t = \mu + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q} \qquad (1)$$

Where, y_t is the actual value at time period t. ε_t is the random error at time period t, ϕ_i (i = 1; 2…..p) and θ_j (j= 0; 1; 2 …q) are model parameters. p and q are integers and frequently mentioned as orders of the model. Random errors ε_t, are assumed to be identically and independently distributed with a mean of zero & a constant variance of σ^2. ARIMA analysis is done on the preprocessed dataset. The predicted values and errors for the individuals are calculated through ARIMA analysis.

These are all happens to study the single attribute. The proposed methodology analysis for N number of

attributes and also detects the p,d, and q values for each chunk of data set separately based on the nature of the value in each attribute. The error rate in the value play as supporting tool to decide the rank of index.

1.2 Time Series Analysis Error Rate Method

The commonly used forecast measures for determining the error rate are

1. Mean Square Error (MSE)
2. Root Mean Squar Error (RMSE)

Mean Squared Error is a measure of the quality of an estimator. In statistics the mean square error (MSE) or root mean squared error (RMSE) of an estimator (of a procedure for estimating an unobserved quantity). The MSE measures the average of the squares of the errors or deviations—that is, the difference between the estimator and what is estimated. This is evaluated using the equation 2.

$$MSE = 1/n \sum_{k=0}^{n} (Yk - Yk - 1)^2 \qquad (2)$$

The RMSE is frequently used measurement to identify the deviation among the values. It is an average of squared error measured using the equation (3).

$$\text{RMSE} = \frac{\sqrt{\sum_{k=0}^{n} (Yk - Y\,k-1)^2}}{\text{n}} \qquad (3)$$

2. Literature Review

Clusters of same data set may differ in dimension. One cluster may have an object with similar data in some dimension. Other cluster may have object with similar data with other set of dimension. But there is need of comparing these differential clusters[1]. Some of the projected space clustering filter less important dimension from the chosen set of attributes[2, 3]. Treat that as less in cluster formation and eliminate those dimensions. The clustering done with the remaining data set. The dimension removed as may influence as predicator in classification in some domain like weather and stock market[4,5].

So the attributes are major inducer of cluster accuracy. There are some clustering algorithms where the attribute plays major role in clustering. The algorithm CACTUS starts from initial core attribute which is not allowed to appear on another cluster[4]. The algorithm COOLCAT use entropy to calculate the closeness of clusters[6]. It needs the K input to select centroid of clusters. The CLICK used graph based partition of data by applying weighted attributes[6]. The attributes are vertex of graph the edge between the vertexes weighted to find the proximity of the vertexes. This will not handled high dimension data. The algorithm PROCAD projects the weight based attributes and rejects the low rank attributes and this works only for categorical data set alone[7]. The method proposed in this paper find the attributes by comparing pair of ratio among attributes and form a rank matrix, from which attributes are ranked. These variable selection allows to add the new variable and also decides the number of clusters.

In this model they introduced a flexible lag structure, this structure is the extension of ARCH process [9]. GARCH process stands more similarly to the extension of standard time series AR process to the general ARMA process are developed, and represents the variance of the error term as a function of its auto regressive terms, thereby allowing a more mean representation of the time series. TAR techniques are applicable in the financial time-series modeling, prices are deviated from where deviations of prices from balanced values depending on the discrete transaction costs and the market regulators rules based on the threshold values of variables [10]. The main finance application in modeling the difference in prices of equal resources in the presence of transaction costs.

The transformation of attribute play vital role in cluster and classification process. The rank based transformation is the achievable scenario for the best outcome in mining process.

3. eARIMA Model with MPSKM Algorithm

The work flow of the proposed model visualized in Fig 1. The model star working from data split. The seasonal based study yields better result when compare to handling the entire time series data as whole. The seasonal variation could be identified easily. The preprocessing fill the missing data with the average of the respective attribute. Even in the preprocessing the average are taken from the near early and near future values of the attribute as because the early and later values simulate the previous and next season respectively and the seasonal changes are also gradual. The internal

season may have sudden variation that to be handled as an outline or an extream event.

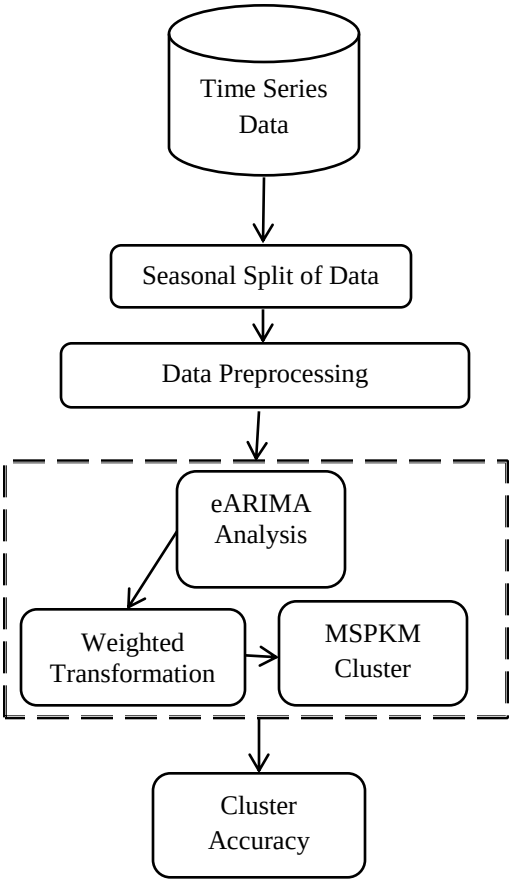

Figure 1 : eARIMA model with MPSKM Algorithm

The preprocessed data with entire set of attribute are taken for the error analysis using eARIMA method. It do the analysis for the entire set of attribute which backup the relation or interdependency of the values. The model also find the ARIMA parameters separately for each season. So the [p,q,d] varies according to the seasonal reading of attribute. The error prediction for each attribute transforms the data set by applying weighted index. The low error attribute allowed with high rank. It support to reduce the iteration of cluster and also to improve the cluster accuracy. The transformed data set are clustered to label using MPSKM a modified version of traditional K-Mean algorithm which decided the K value equal to the number of attributes in data set.

4. Result and Discussion

The data used is the daily weather time series data for the period from January-1982 to July- 2014. Dataset compromise of 11899 data points. The attribute includes Temperature (Min, Max), Wind speed, Rain Fall, Humidity1,Humidity 2 (RH1, RH2), Radiation Heat (RH), Solar Scale(SS), Evaporation (EVP) . All these attributes are studied using eARIMA to get the individual error rate for each season of data. The seasonal study always yield the better result [8]. The eARMA parameters for each season are identified separately, do to which these parameters takes suitable values. The moving average ranges to the maximum of 3, so these parameters take the probabilistic ranges from [0,0,0] to [1,2,3] respectively for p,q and d. The lower error rate attribute applied with high ranking and the higher error attribute applied with lower rank. These study report are shown in the Fig 2, 3,4 for the year of data 982,1983 and 2014.

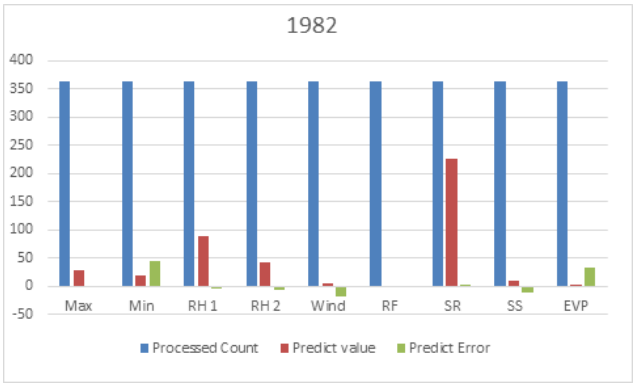

Figure 2 : eARIMA Error Prediction - 1982

The eARIMA suggest the rank of attribute based on the error rate. The transformed data set of each season are clustered using MPSKM algorithm. The number of clusters are equal to the number of

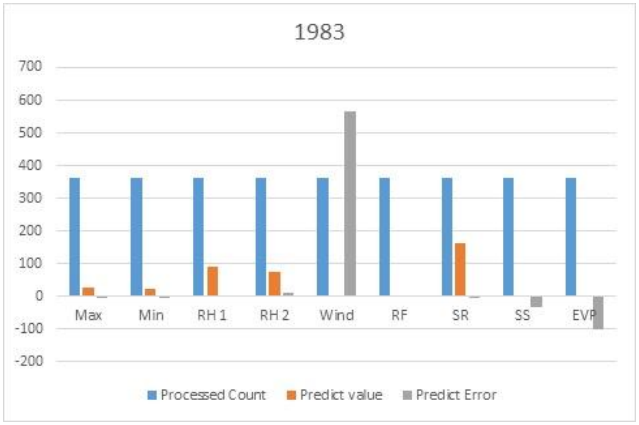

Figure 3 : eARIMA Error Prediction - 1983

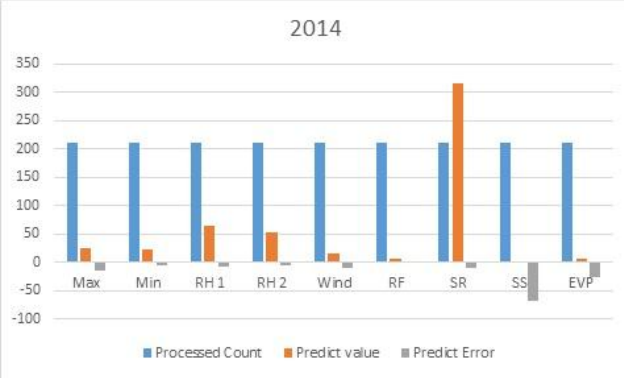

Figure 4 : eARIMA Error Prediction - 2014

Table 1 – Mean Square Error

Season	Max	Min	RH 1	RH 2	Wind	RF	SR	SS	EVP
1982									
Season 1	1.0156	1.0594	0.407	0.8926	0.3687	0.0005	3.0803	0.7053	0.5854
Season 2	0.9777	0.5756	0.6708	1.9985	2.0568	0.1642	10.1393	3.3023	0.6354
Season 3	4.2205	0.8809	0.6716	3.2457	7.2186	0.725	20.203	6.9833	1.3792
Season 4	2.2038	1.4118	0.7344	4.2408	7.7904	2.421	26.8817	10.5076	2.237
1983									
Season 1	0.6127	0.7745	0.6787	0.7479	0.4108	0	2.1699	1.1859	0.3469
Season 2	0.4394	0.3711	0.9571	1.7549	1.8357	0.0877	4.2354	3.1105	0.7213
Season 3	1.0677	0.4806	1.2134	2.4548	5.9541	0.5802	10.4278	6.7687	1.7852
Season 4	2.7158	1.8084	1.0676	3.943	8.4564	1.8994	23.6385	11.0282	2.7486
2014									
Season 1	0.3719	1.061	0.3509	0.6085	0.5983	0.0002	0.9024	1.4132	0.0055
Season 2	0.807	1.7092	0.6164	1.0055	1.8549	0.0007	1.163	3.5914	0.0126
Season 3	0.9906	0.9515	0.918	1.4828	2.3361	0.0464	1.8229	4.7807	0.0189
Season 4	3.9134	1.8376	1.1009	3.17	5.6508	0.4474	3.7594	7.621	0.0266

Table 2 – Root Mean Square Error

Season	Max	Min	RH 1	RH 2	Wind	RF	SR	SS	EVP
1982									
Season 1	1.0078	1.0293	0.638	0.9448	0.6072	0.0232	1.7551	0.8398	0.7651
Season 2	0.9888	0.7587	0.819	1.4137	1.4342	0.4053	3.1842	1.8172	0.7971
Season 3	2.0544	0.9385	0.8195	1.8016	2.6868	0.8514	4.4948	2.6426	1.1744
Season 4	1.4845	1.1882	0.857	2.0593	2.7911	1.556	5.1848	3.2415	1.4957
1983									
Season 1	0.7828	0.88	0.8239	0.8648	0.641	0	1.4731	1.089	0.589
Season 2	0.6628	0.6092	0.9783	1.3247	1.3549	0.2962	2.058	1.7637	0.8493
Season 3	1.0333	0.6933	1.1015	1.5668	2.4401	0.7617	3.2292	2.6017	1.3361
Season 4	1.648	1.3448	1.0332	1.9857	2.908	1.3782	4.8619	3.3209	1.6579
2014									

Season 1	0.6098	1.03	0.5924	0.78	0.7735	0.0136	0.9499	1.1888	0.0742
Season 2	0.8983	1.3074	0.7851	1.0028	1.3619	0.0272	1.0784	1.8951	0.1123
Season 3	0.9953	0.9754	0.9581	1.2177	1.5284	0.2155	1.3501	2.1865	0.1374
Season 4	1.9782	1.3556	1.0492	1.7805	2.3771	0.6689	1.9389	2.7606	0.1631

error rate are tabulated in Table 1 and Two the same of MSE visualized in Fig 5,6,7 respectively.

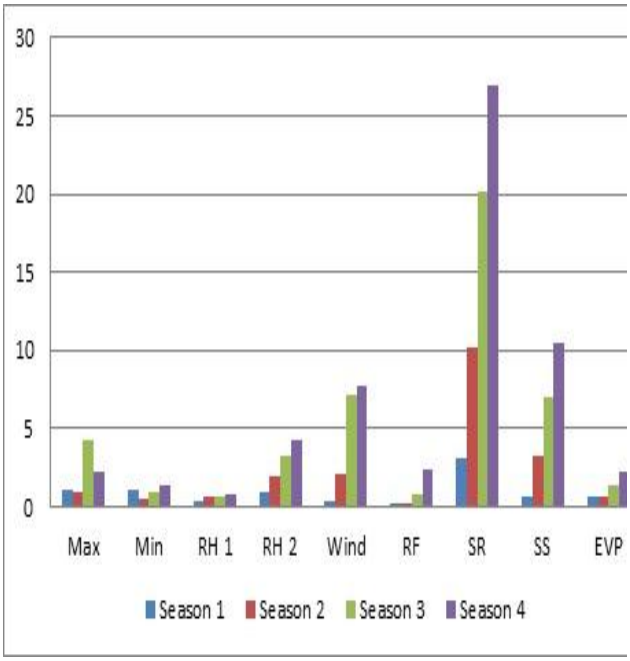

Figure 5 – MSE of 1982

attributes in the data set. In the experimental the value of K=9. The first object is taken as the center of each cluster. Individual deviation of the object from the center has evaluated to decide the cluster of object. The similarity of the objects are identified using MSE and RMSE. These method reduces the

number of iteration in cluster finalization and also gives the improvised cluster accuracy when compare to the MPSC(Modified Projected Space Cluster) which take the automated K value alone and manual threshold to decide the MSE of finding object similarity. The cluster accuracy of three years of dataset are measured using MSE and RMSE. The

The attribute with minimal or negative predict error in eARIMA produces almost 99% of accuracy in cluster accuracy for which the cluster based on the same attribute. Other clusters also produces 95% and above accuracy. This accuracy considerably improvised rate compare to MSPC method.

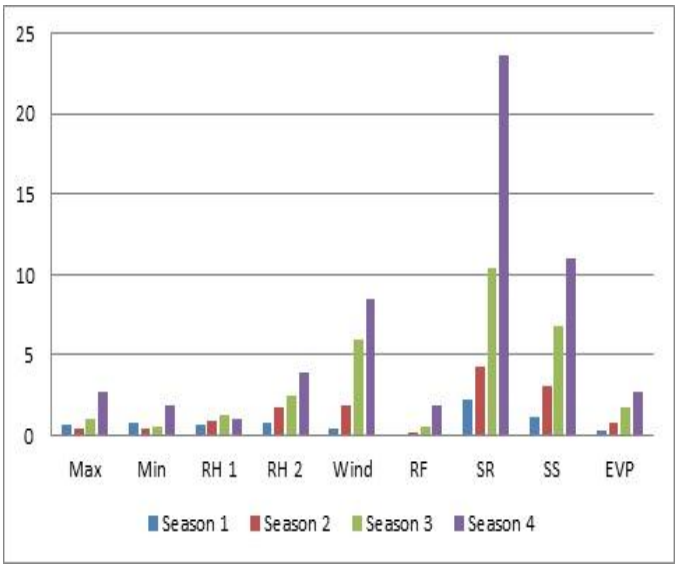

Figure 6 – MSE of 1983

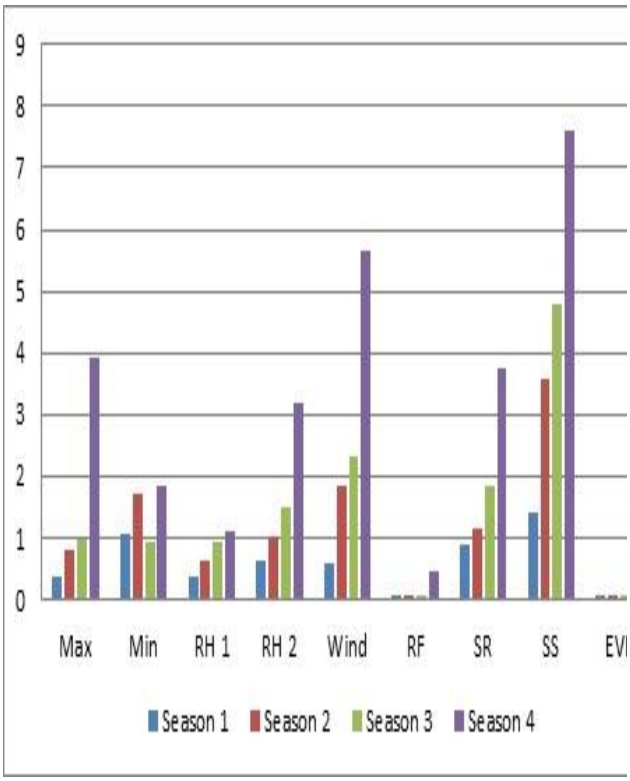

Figure 7 – MSE of 2014

Conclusion

Time series analysis is the trendy analysis to react to the environment. The proposed eARIMA model handles all the attributes in a data set for ranking with weighted. The high error attribute with low weight is an added advantage to find the lower outline pattern in the event of data. The MPSKM clustering with eARIMA produces the considerable accuracy improvement. The future work can enhance the method to do the prediction based on the ensemble produced and also the improvisation in ensemble by applying the dynamic clustering to reduce the number of clusters.

References

1. Sung-Soo Kim, Variable Selection and Outlier Detection for Automated K-means Clustering, Communications for Statistical Applications and Methods, Vol 22, No 1, 2015, pp 55–67.
2. Zaki Mohammed J, Peters M, Assent I, Seidl T, CLICKS: an effective algorithm for mining subspace clusters in categorical datasets., Data & Knowledge Engineering, Vol 60, No 1, 2007, pp 51–70.
3. Cesario E, Manco G, Ortale R, Top-down parameter-free clustering of high-dimensional categorical Data, Knowledge and Data Engineering, IEEE Transactions, Vol 19, No 12, 2007, pp 1607–1624.
4. Amardeep Kaur, AmitavaDatta, A novel algorithm for fast and scalable subspace clustering of high-dimensional data, Journal of Big Data, Vol 2, No 17, 2015, pp 1-24.
5. KavitaThawkar, SnehalGolait, RushiLongadge, A Framework for an Outlier Pattern Detection in Weather Forecasting, IJCSMC, Vol 3, No 5, 2014, pp 348 – 358
6. Barbara D, Li Y, Couto J ,COOLCAT: an entropy-based algorithm for categorical clustering, Proceedings of the 11th ACM international conference on information and knowledge management (CIKM'02), Vol 1, No1, 2002, pp 582–589.
7. M. Bouguessa, Clustering categorical data in projected spaces, Data Mining and Knowledge Discovery, Vol 29, No 1, 2015, pp 3–38.
8. S. Gokila, K. Anandakumar, A.Bharathi, Modified Projected Space Clustering Model on Weather Data to Predict Climate of Next Season, Indian Journal of Science and Technology,Vol 8, No 14, 2015, pp 1-5
9. Bollerslev, T. Generalized autoregressive conditional heteroskedasticity. Journal of econometrics, Vol 31, No 3, 1986, pp 307-327.
10. Yadav, P. K., Pope, P. F., & Paudyal, K.. Threshold autoregressive modeling in finance: The price differences of equivalent assets. Mathematical Finance, Vol 4, No 2, 1994, pp 205-221.