Prediction, Classification and Recommendation in e-Health via Contextual Partitioning

Muhammad Anjum Qureshi
Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey
qureshi@ee.bilkent.edu.tr

***Abstract*—In this paper, we propose a multipurpose contextual partitioning based estimation algorithm. Exploiting the similarities between contexts (side information: such as age, Gender etc.,) related to patient data in healthcare repository or database, multidimensional spheres are generated over Euclidean space. Then, conditional first and second order characteristics are predicted using sample-based mean and covariance. These conditional statistics of particular patient data subset (sphere) serve the following purposes: i) Prediction for missing values (conditional mean), ii) Partitioned principal components for better classification (conditional covariance) and iii) Recommendation for medical Test or physician (conditional covariance). The proposed approach uniformly partitions the context space into spheres, and then, for each sphere estimates the conditional mean and covariance using only the data (excluding the context data) in the selected sphere. Hence, providing three in one solution i.e., Prediction, Classification and Recommendation for healthcare data using conditional probabilistic characteristics. The overall error is decomposed into estimation and approximation errors. In a particular sphere, estimation error is dependent on the number of instances, while approximation error is dependent on the dissimilarity of instances.**

***Index Terms*—Prediction, classification, recommendation, uniform partition.**

I. Introduction

Healthcare informatics is considered to be the most important application in semantic computing [1], [2]. In recent time, healthcare solution providers are adopting the electronic health records (EHRs) for information mining. These huge datasets are attractive to data mining expert to analysis this big data and provide decision making for patient health. The context or conditioning variable which can be patient health record is used to extract the relevant knowledge from the related database, and solution is provided using selected subset of the database for most of the decision providing systems.

In this paper, we propose partition-based estimations where conditional statistics are estimated by partitioning the database based on context, and estimating the parameters based only on the data samples excluding the context samples, for the selected sphere in the partition. The overall error of the conditional mean and covariance is decomposed into two parts: (i) approximation error, which is proportional to the diameter/radius of the sphere in the set for which estimation is performed, and (ii) estimation error, which is related to the finite sample size of the sphere in the set for which estimation is performed. There is a trade-off between these two errors, where increasing the number of instances that fall within a particular sphere (which is equivalent to increasing the length of the set) decreases the estimation error but increases the approximation error in that sphere. The proposed method serves multipurpose in Health Recommendation Systems (HRS), where its estimation parameters are used to predict the missing values in the dataset, partitioned principle components providing better classification, and recommendation for the physician or medical test based on the analysis of similar cases. In the proposed algorithm, arriving patient data is treated as the context vector, the health record database is the feature space. The conditional structure provides the relevancy between the patient data and available health record system.

The heteroscedastic regression models are considered to be widely implemented algorithm for estimation of the conditional mean and covariance. A variant of a bilinear model, Autoregressive Conditional Heteroscedasticity (ARCH) was proposed to estimate the conditional covariance matrix, followed by Generalized Autoregressive Conditional Heteroscedasticity (GARCH) [3], [4]. These ARCH method explicitly recognizes the difference between unconditional and conditional variance, and represents the conditional variance as function of past errors, whereas the GARCH method provides active learning mechanism to ARCH method [5]. Considering the estimation of $\sigma^2(x)$ as a non-parametric regression, different kernal based approaches were proposed including, residual-based estimator by the local linear technique [6], double-autoregressive model [7]. The main drawback of the proposed methods is that the resultant estimated covariance matrix is not always positive. The variations were then proposed to provide always positive estimate by introducing log transformations [8], [9].

In nonparametric estimation, natural way to target the local properties is to partition the data sapce into subsets, and estimate the statistics of each subset independently. The simplest method partitions the data space into rectangle or cubes based on the length and size of the dataset. The resultanat estiamte is consistent regardless of the data distribution [10]. These data-dependent partition schemes are implemented in real world applications and superiority is shown over fix sequence of partitions [11], [12]. Given i.i.d random variables, it is provided in the literature that for empirical probability measure using m samples, the estimation approaches to the true value fo $m \to \infty$ [10]. The basis algorithm first partitions the data space into subsets such that each set contains equal number of samples. The empirical measure is calculated using

978-1-6654-3649-6/21/$31.00 ©2021 IEEE

2021 29th Signal Processing and Communications Applications Conference (SIU) | 978-1-6654-3649-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/SIU53274.2021.9477921

Authorized licensed use limited to: ULAKBIM UASL - Bilkent University. Downloaded on February 01,2022 at 11:00:13 UTC from IEEE Xplore. Restrictions apply.

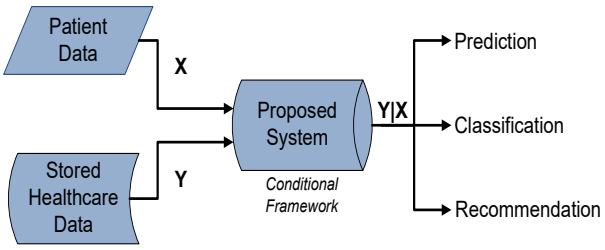

Fig. 1. The Proposed Multipurpose System

the samples that fall inside particular subset. In conventional partitioning methods [13], [14], variables responsible for partitioning the data space are also used to estimate the local statistics.

In our proposed method, the data is separated into two spaces, one is the context space and other is hidden feature space. The partitioning is carried out based only on the context space. The hidden feature space is not used in this step. Once the sets are formed, the hidden feature space inside any selected set is used to estimate the conditional statistics using maximum likelihood method. The samples selected in the particular set in the partition are hence dependent on the selected context.

II. Problem Formulation

The system model is shown in Fig. 1. Patient health record denoted by X is arrived at the context input of the proposed system, the system searches the relevant data in the patient record database denoted by Y, and provides the prediction, classification and recommendation via utilizing the extracted information. This information represents the conditional statistics denoted by $Y|X$, where different tasks are dependent on first order or second order statistics.

Let $\boldsymbol{Z} = (\boldsymbol{Y}, \boldsymbol{X})$, where $\boldsymbol{X} \in \mathcal{X} := \mathbb{R}^q$ is context vector and $\boldsymbol{Y} \in \mathcal{Y} := \mathbb{R}^p$ is feature vector. We assume that each row of $\boldsymbol{Z}$ is sampled from an independent and identically distributed (i.i.d.) unknown stochastic process, whose *probability density function* (pdf) is denoted by $f_{\boldsymbol{Z}}(\boldsymbol{z})$. For such process, we assume that the conditional pdf of $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$ also exists, and is denoted by $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$. Furthermore, $\Sigma_{\boldsymbol{Y}|\boldsymbol{X}}$ denotes the conditional covariance matrix of the feature vector $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$, and $\mu_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}$ denotes the conditional mean of the feature vector $\boldsymbol{Y}$ given $\mathbf{X} = \boldsymbol{x}$.

We assume that the logged dataset is composed of N samples from this process, which is denoted by $\boldsymbol{Z} := [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N]^T$, where $\boldsymbol{z}_n = (\boldsymbol{x}_n, \boldsymbol{y}_n)$ denotes the nth sample. Our goal is to estimate $\Sigma_{\boldsymbol{Y}|\boldsymbol{X}}$ and $\mu_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}$ for all context vectors in context space $\mathcal{X}$, using selected feature vectors from feature space $\mathcal{Y}$. We say that the estimated conditional covariance matrix $\hat{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}}$ is (ϵ, δ) optimal if

$$\Pr(||\hat{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}|\boldsymbol{X}}|| \leq \epsilon ||\Sigma_{\boldsymbol{Y}|\boldsymbol{X}}||) > 1 - \delta$$

and the estimated conditional mean vector $\hat{\mu}_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}$ is (ϵ, δ) optimal if

$$\Pr(||\hat{\mu}_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}} - \mu_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}|| \leq \epsilon ||\mu_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}||) > 1 - \delta$$

where $||(\cdot)|| = ||(\cdot)||_2$ denotes the l_2-norm for vectors and induced spectral norm for matrices. Let g be a vector and Σ be a covariance matrix having $\Sigma^T = \Sigma$ with eigenvalue denoted by λ and maximum eigenvalue by λ_{max}, then spectral norm for Σ is calculated as

$$||\Sigma|| = \sup_{g \neq 0} \frac{||\Sigma g||_2}{||g||_2} = \sqrt{\lambda_{max}(\Sigma^T \Sigma)} = \lambda_{max}(\Sigma) \quad (1)$$

Definition 1. *($\mathcal{S}$): The context subspace $\mathcal{S}$ is the bounded subset of $\mathcal{X}$ using the characteristics of eigenvalues. The diagonal elements of context covariance matrix provides the variances in each axis, which is not guaranteed to be the direction of maximum variance. The maximum eigenvalue of this covariance matrix denoted by λ_{max} provides the direction of maximum variance.*

Two definitions associated with bounded spaces are:
covering number $\mathcal{N}$: Any set of balls with radius D covering a space $\mathcal{S}$ is called *D-cover* of $\mathcal{S}$. The set of their centers is *D-net*. The cardinality of the smallest such *D-net* is called covering number.
packing number $\mathcal{M}$: A subset of elements lies in $\mathcal{S}$ is said to be *D-separated* if distance between all distinct elements pair is at least D. The count of maximum elements that can be *D-separated* is called packing number.

III. Prediction, Classification, Recommendation via Uniform Partitioning

In this section we propose an algorithm that estimates $\Sigma_{\boldsymbol{Y}|\boldsymbol{X}}$ and $\mu_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}$ from the logged dataset by uniformly partitioning the dataset according to the similarities between the contexts.

A. Algorithm

Let $\mathcal{P} := \{s_1, s_2, \ldots, s_{m_D}\}$ denotes a D-cover of context subspace $\mathcal{S} \subset \mathcal{X}$ containing m_D number of equally sized ($D :=$ largest distance of context from center, radius of the sphere) context spheres. The centers of these spheres denoted by $c_s, \forall s \in \mathcal{P}$ are not allowed to overlap due to the fact that these points are representative points of contextual spheres. Set of these centers is called D-separated set of $\mathcal{S}$ and is denoted as $C_{\mathcal{S}} := \{c_{s_1}, \ldots, c_{s_{m_D}}\}$, as the distance between centers is at least D.

The center of first sphere is the context closest to mean vector of the context, as it provides maximum point on the density function, and covers maximum possible area with fix D. The context mean and covariance are estimates using sample based method. After forming first sphere around mean vector, center of the second sphere is the closest context outside first sphere, hence allowing some elements to overlap. Our goal is to estimate sample based conditional mean and covariance for every contextual sphere, under the criterion that samples of feature vector that fall inside selected sphere are used. The

Authorized licensed use limited to: ULAKBIM UASL - Bilkent University. Downloaded on February 01,2022 at 11:00:13 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1 ESUP

Input: *N, p, q, D*
Initialize: *Create partition $\mathcal{P}$of logged data into m_D spheres each with size D containing X, Y*
while $n \geq 1$
  Find the spheres in $\mathcal{P}$ that x_n belongs to, i.e., s_n
  $\bar{\mathcal{P}}$ denotes the set of these selected spheres.
  $\hat{\mu}_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}_n} = \arg\max_{N_s}\{\hat{\mu}_{\boldsymbol{Y}|s\in\bar{\mathcal{P}}}(N_s)\}$
  $\hat{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}_n} = \arg\max_{N_s}\{\hat{\Sigma}_{\boldsymbol{Y}|s\in\mathcal{P}^*}(N_s)\}$
  $n = n + 1$
end while

Algorithm 2 Create Partition

Estimate: *Context Characteristics*
$\hat{\mu}_{\boldsymbol{X}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i)$
$\hat{\Sigma}_{\boldsymbol{X}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \hat{\mu}_{\boldsymbol{X}})^T(\boldsymbol{x}_i - \hat{\mu}_{\boldsymbol{X}})$
Initialize: $j = 1, s_j = \{\}, \{\bar{\mathcal{S}}\} = \{\mathcal{S}\}$
$c_{s_j} = \arg\min_{\boldsymbol{x}_j}\{||\hat{\mu}_{\boldsymbol{X}} - \boldsymbol{x}_j||\}$
while $\{c_{s_j} \neq \emptyset\}$
  $s_j = \{c_{s_j}\} \cup \{\boldsymbol{x}_i \mid ||c_{s_j} - \boldsymbol{x}_i|| \leq D\}, \forall \boldsymbol{x}_i \in \bar{\mathcal{S}}$
  $\hat{\mu}_{\boldsymbol{Y}|\boldsymbol{s_j}} = \frac{1}{N_{s_j}}\sum_{i\in I(s_j)}(\boldsymbol{y}_i)$
  $\hat{\Sigma}_{\boldsymbol{Y}|\boldsymbol{s_j}} = \frac{1}{N_{s_j}}\sum_{i\in I(s_j)}(\boldsymbol{y}_i - \hat{\mu}_{\boldsymbol{Y}|\boldsymbol{s_j}})^T(\boldsymbol{y}_i - \hat{\mu}_{\boldsymbol{Y}|\boldsymbol{s_j}})$
  $j = j + 1$
  $\bar{\mathcal{S}} \leftarrow \{\bar{\mathcal{S}}\} - \{s_j\}$
  $c_{s_j} = \arg\min_{\boldsymbol{x}_j\in\bar{S}}\{||c_{s_{j-1}} - \boldsymbol{x}_j||\}$
end while

sphere in the partition to which arrived context belongs is selected (sphere with maximum cardinality is selected in case arrived context belongs to two or more spheres, as for fix D maximum cardinality sphere provides the best estimate) and estimation for conditional mean and covariance is provided. Cardinality of sphere s is denoted by N_s. This algorithm is named as *Estimator Selection with Uniform Partitioning* (ESUP), and its pseudo-code is given in Algorithm 1 followed by pseudo-code of contextual partition in Algorithm 2. The second sphere is created around the closest context vector that is outside of first sphere based on euclidean distance, and the process continues until all the significant subspace $\mathcal{S}$ is covered.

Let $I(s) := \{1 \leq i \leq N : \boldsymbol{x}_i \in s\}$, denotes the indices of feature instances for which $\boldsymbol{x}_i \in s$, then the estimates are calculated as,

$$\hat{\mu}_{\boldsymbol{Y}|s} = \frac{1}{N_s}\sum_{i\in I(s)} \boldsymbol{y}_i$$

and

$$\hat{\Sigma}_{\boldsymbol{Y}|s} = \frac{1}{N_s}\sum_{i\in I(s)} (\boldsymbol{y}_i - \hat{\mu}_{\boldsymbol{Y}|s})^T(\boldsymbol{y}_i - \hat{\mu}_{\boldsymbol{Y}|s}).$$

Assumption 1. *The number of samples N are assumed to be much more larger than dimensions of context and feature vectors $N >> p, q$, providing at least one sphere exists with cardinality $N_s > p$.*

Assumption 2. *There exists $L_1, L_2 > 0$ such that for all $\boldsymbol{x}_i, \boldsymbol{x}_j \in s, \forall s \in \mathcal{P}$, we have $||\mu_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x_i}} - \mu_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x_j}}|| \leq L_1||\boldsymbol{x_i} - \boldsymbol{x_j}||$ for conditional mean, and $||\Sigma_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x_i}} - \Sigma_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x_j}}|| \leq L_2||\boldsymbol{x_i} - \boldsymbol{x_j}||$, for conditional covariance.*

The above assumption indicates that the estimation quality is same for similar contexts belonging to a particular sphere.

IV. Performance Evaluation

This section describes the performance evaluation criteria and benchmark algorithms to compare with the proposed algorithm.

A. Dateset Description

We perform the experiments over the Thyroid disease dataset available in UCI (Center of Machine Learning and Intelligent Systems, University of California) Machine Learning Repository. There are 29 features available in the dataset, 7 features are continuous and 22 features are binary out of total 29 features [15], [16]. In training dataset, there are total of 2800 instances (cases), whereas 972 instances are provided for testing. We use *classification and regression trees* (CART) decision trees for classification purposes [17].

B. Experiment 1: Prediction/Missing Values (Fig. 2)

The estimated conditional mean by the ESUP serves as the missing value estimator or prediction. The partitioned spheres are created based on the similarities between context vectors, and hence the mean value of the feature vector provides the best average behavior of the feature. The feature with missing values are not used in the partitioning process, and are used as the arrivals to the system. For each new arrival, the algorithm provides the conditional mean as the estimated value for the missing value in the feature vector.

An experiment is performed for the prediction of the missing values. We select the 'allrep' dataset provided in the repository, and select referral source feature as the context. We compare the predictions provided by i) standard mean method which calculates the overall mean, ii) standard median method which calculates the overall median, iii) ESUP which utilizes the context information to predict missing value via conditional statistics. We randomly select the 70% of the available training dataset to learn the standard mean, standard median and conditional statistics (for ESUP), and then, predict some of the values of features in the remaining 30% data by removing the available values. The results in terms of the normalized norm squared error (l_2-norm is calculated between the predicted values and the original values removed from the dataset, squared and then normalized in $(0, 1)$) are provided in Fig. 2. The presented results are obtained by averaging over 100 independent runs. It is shown that utilizing the conditional mean provides lower error when compared to the standard mean and median estimation.

Authorized licensed use limited to: ULAKBIM UASL - Bilkent University. Downloaded on February 01,2022 at 11:00:13 UTC from IEEE Xplore. Restrictions apply.

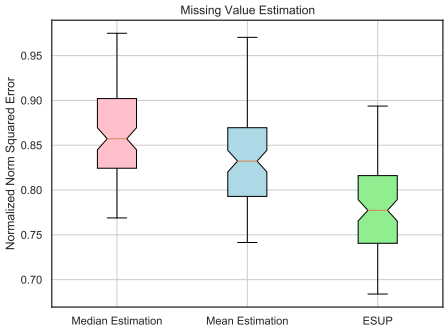

Fig. 2. Missing value prediction experiment

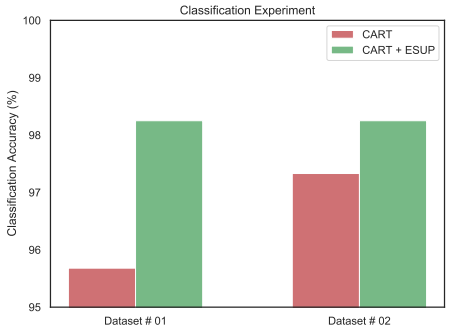

Fig. 3. Classification/Recommendation experiment

C. Experiment 2: Classification/Recommendation (Fig. 3)

The estimated conditional covariance by the ESUP provides the basis for maximum variance analysis and recommending the maximum valuable medical test (i.e., feature). The partitioned spheres are created based on the similarities between context vectors, and hence the eigenvalues and vector of conditional covariance matrix of the feature vector provides the maximum variance direction. The feature space contains independent principal components for each partition sphere in contrast to single principal component structure of overall data. It provides better relationship between the reduced (recommended) feature data to the classification.

An experiment is performed for the classification of the thyroid disease in two of the dataset provided in the repository named 'allhypo' and 'allrep'. We compared the predictions provided by i) standard covariance estimation method which estimates the overall covariance matrix and select best 18 features, and then decision tree algorithm is applied for classification ii) ESUP which utilizes the context information (referral source feature) to estimate the conditional covariance, and based on conditional statistics recommends best 18 features given that context, and then, decision tree algorithm is applied for classification over reduced feature space. We also normalize the features in $(0, 1)$ before applying CART decision trees. We train the decision tree from the training dataset, and the trained decision tree is used over the test data for classification. The results in terms of classification accuracy are provided in Fig. 3. It is shown that utilizing the conditional covariance provides better classification accuracy and recommends better reduced feature space.

V. Conclusion

This study provides the estimation of conditional mean and covaraince in computationally efficient way (e.g., without the matrix inversion in case of Gaussian distributed data). The candidate applications of the proposed conditional framework in e-Health system are provided, which includes prediction, classification and recommendation tasks. An algorithm based on contextual partitioning is proposed with the overall error decomposed in the approximation error and the estimation error. The Uniform partition method provides the set of equally sized spheres, and the maximum cardinality sphere among these spheres is the one with lowest error. Furthermore, the estimated conditional covaraince matrix is always semi-positive definite. The experimental results are provided for the thyroid disease diagnosis.

References

[1] P. C.-Y. Sheu, H. Yu, C. Ramamoorthy, A. K. Joshi, and L. A. Zadeh, *Semantic computing*. John Wiley & Sons, 2011.

[2] C. C. Wang, D. A. Hecht, P. C.-Y. Sheu, and J. J. Tsai, "Semantic computing and drug discovery-a preliminary report," in *Proc. 7th IEEE Int. Conf. Semantic Computing (ICSC)*, pp. 453–458, 2013.

[3] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *J. Econometrics*, vol. 31, no. 3, pp. 307–327, 1986.

[4] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica: J. Econometric Society*, pp. 987–1007, 1982.

[5] D. B. Nelson, "Filtering and forecasting with misspecified ARCH models I: Getting the right variance with the wrong model," *J. Econometrics*, vol. 52, no. 1-2, pp. 61–90, 1992.

[6] J. Fan and Q. Yao, "Efficient estimation of conditional variance functions in stochastic regression," *Biometrika*, pp. 645–660, 1998.

[7] S. Ling, "Estimation and testing stationarity for double-autoregressive models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 66, no. 1, pp. 63–78, 2004.

[8] L.-H. Chen, M.-Y. Cheng, and L. Peng, "Conditional variance estimation in heteroscedastic regression models," *J. Statistical Planning and Inference*, vol. 139, no. 2, pp. 236–245, 2009.

[9] K. Yu and M. Jones, "Likelihood-based local linear estimation of the conditional variance function," *J. American Statistical Association*, vol. 99, no. 465, pp. 139–144, 2004.

[10] G. Lugosi, A. Nobel, *et al.*, "Consistency of data-driven histogram methods for density estimation and classification," *The Annals of Statistics*, vol. 24, no. 2, pp. 687–706, 1996.

[11] C. Tekin and M. van der Schaar, "Active learning in context-driven stream mining with an application to image mining," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3666–3679, 2015.

[12] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3064–3074, 2005.

[13] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: A conditional inference framework," *J. Computational and Graphical Statistics*, vol. 15, no. 3, pp. 651–674, 2006.

[14] C. Tekin and M. van der Schaar, "Distributed online learning via cooperative contextual bandits," *IEEE Trans. Signal Process.*, vol. 63, no. 14, pp. 3700–3714, 2015.

[15] M. A. Qureshi and K. Eksioglu, "Expert advice ensemble for thyroid disease diagnosis," in *Proc. 25th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2017.

[16] K. Ekşioğlu, M. A. Qureshi, and C. Tekin, "Online classification with contextual exponential weights for disease diagnostics," in *Proc. 25th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2017.

[17] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

Authorized licensed use limited to: ULAKBIM UASL - Bilkent University. Downloaded on February 01,2022 at 11:00:13 UTC from IEEE Xplore. Restrictions apply.