International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:12, No:8, 2018

Distances over Incomplete Diabetes and Breast Cancer Data Based on Bhattacharyya Distance

Loai AbdAllah, Mahmoud Kaiyal

***Abstract*—**Missing values in real-world datasets are a common problem. Many algorithms were developed to deal with this problem, most of them replace the missing values with a fixed value that was computed based on the observed values. In our work, we used a distance function based on Bhattacharyya distance to measure the distance between objects with missing values. Bhattacharyya distance, which measures the similarity of two probability distributions. The proposed distance distinguishes between known and unknown values. Where the distance between two known values is the Mahalanobis distance. When, on the other hand, one of them is missing the distance is computed based on the distribution of the known values, for the coordinate that contains the missing value. This method was integrated with Wikaya, a digital health company developing a platform that helps to improve prevention of chronic diseases such as diabetes and cancer. In order for Wikaya's recommendation system to work distance between users need to be measured. Since there are missing values in the collected data, there is a need to develop a distance function distances between incomplete users profiles. To evaluate the accuracy of the proposed distance function in reflecting the actual similarity between different objects, when some of them contain missing values, we integrated it within the framework of k nearest neighbors (kNN) classifier, since its computation is based only on the similarity between objects. To validate this, we ran the algorithm over diabetes and breast cancer datasets, standard benchmark datasets from the UCI repository. Our experiments show that kNN classifier using our proposed distance function outperforms the kNN using other existing methods.

***Keywords*—**Missing values, distance metric, Bhattacharyya distance.

I. Introduction

MANY real-world datasets suffer from the problem of missing values. There are many serious data quality problems in health datasets such as: missing, redundant, inconsistent, outliers and noisy data. Missing values can be caused by human errors, system generated errors, equipment failure, and so on. Based on the research of Cabena [3], about 20% of the effort is spent trying to solve the problem and on figuring out the data, 60% of the effort is spent on data preparation and feature extraction and another about 20% on data analysis. We were introduced to this problem by a data we received from Wikaya ltd, an artificial intelligence platform that helps improve prevention to chronic diseases. We do that by providing the Prevention Score giving a clear indication to the level of prevention efforts. The Score is calculated based on clinical algorithms licensed from Washington University. The platform define users profiles by collecting data from the users, their mobile phones, and wearables and, in phase II, through integration with EMR. Therefore many values will be assigned as missing values.

In this research we use a distance function over incomplete users profile based on Bhattacharyya distance, where some patients do not have full health profiles. Our work is based on the work of Abdallah et al. [1]. Today, the existing mehtods, solve this problem by filling the missing values with a fix value that computed based on the known values.

However, our suggested method is based mainly on Bhattacharyya distance, which measures the similarity of two probability distributions. We distinguished between two cases: (a) complete patients profiles and (2) incomplete profiles. Where the distance between two complete profiles is simply the Mahalanobis distance. When on the other hand there is a missing value of one of the attributes, the distance is computed based on to the distribution of the missing attribute- attribute might be a risk factor or data collected from the smarphones.

To measure the ability of the derived distance function to reflect the actual similarity between different objects when some of their values are missing, we integrated it within the framework of k nearest neighbors (kNN) classifier since its performance is based only on the similarity between objects.

We use standard benchmark data from the UCI repository for both diabetes and breast cancer diseases. Our experiments show that kNN classifier using the developed distance function outperforms the kNN using other methods.

The paper is organized as follows. Previous methods which deal with missing values are reviewed in Section II. An overveiw of Bhattacharyya distance is described in Section III. The proposed distance function using is described in Section IV. Experimental results on numerical diabetes and breast cancer datasets is presented in Section V. Finally, our conclusions are presented in Section VI.

II. Related Work

Several methods have been proposed to deal with missing data.

Based on [4], [7]-[9] there are three main types of missing data:

1) Missing Completely at Random(MCAR): when the missing value is not related to any other sample.
2) Missing at Random(MAR): when the probability that a value is missing may depend on some known values but it does not depend on the other missing values.
3) Not Missing at Random(NMAR): when the probability that a known value is missing depends on the value that would have been observed.

L. AbdAllah is with the Department of Information Systems, The Max Stern Yezreel Valley Academic College, Israel (e-mail: loaia@yvc.ac.il).
M. Kaiyal is with the Reasearch Department, Wikaya Ltd, Israel (e-mail: Mahmoud@wikayamed.com).

There are two basic types of methods to deal with the problem of incomplete datasets. (1) Deletion: methods from this category ignore all the incomplete profiles. These methods may change the distribution of the data by decreasing the volume of the dataset [11]. Moreover, in our case we can not use it because it means to ignore the patients that have missing values in their profiles which is unaccepted in Wikaya case. (2) Imputation: in these methods the missing values were replaced with known value according to statistical computation. Based on these methods we convert then incomplete data to complete data and as a result the exist machine learning algorithms can be run they deal with complete data.

One of the most common approaches in this domain is the Mean Imputation (MI) method that replace each incomplete datapoint with the mean of the data. There are several obvious disadvantages to this method. (a) using a fixed instance to replace all the incomplete instances will change the distribution of the original dataset, (b) ignoring the relationship among attributes will bias the performance of subsequent data mining algorithms. These problems were caused since we replace all the incomplete instance with a fixed one. On the other hand, a variant of this method is to replace the missing values only based on the distribution of the attributes. It means that the algorithm will replace each missing value with the mean of the of its attribute (MA) and the whole instance [10]. And in a case that the values were discrete the missing value will be replaced by the most common (MCA) value in the attribute [6] (i.e., filling the unknown values of the attribute with the value that occurs most often for the same attribute). All those methods ignore the other possible values of the attribute and their distribution and represent the missing value with one value, that is wrong in realworld datasets.

Finally, the k-**Nearest Neighbor Imputation** method [12], [2] estimates the values that should be replaced based on the k nearest neighbors based only on the known values. The main obstacle of this method is the runtime complexity.

III. Bhattacharyya Distance

For completeness we will now give a short overview of the Bhattacharyya Distance and then we will describe how we integrated it within the distance function. A. Bhattacharyya was a statistician who worked in the 1930s at the Indian Statistical Institute. He defined a distance function that measures the similarity/dissimilarity between two given probability distributions.

Consider two univariate probability density functions, f_1, f_2 in the same domain. The Bhattacharyya distance is defined based on the amount of the overlap between two statistical sample as:

$$D_B(f_1, f_2) = -\ln\left(BC(f_1, f_2)\right)$$

where BC is the Bhattacharyya coefficient, which is the amount of overlap between two statistical samples or populations. For discrete probability distributions case the Bhattacharyya coefficient will be:

$$BC(f_1, f_2) = \sum_{x \in X} \sqrt{f_1(x) \cdot f_2(x)},$$

and

$$BC(f_1, f_2) = \int \sqrt{f_1(x) f_2(x)}\mathrm{d}x,$$

for continuous case.

Let $f_1(x), f_2(x)$ be two univariate Gaussian probability density functions as a special case, and assume that $\mu_1 \neq \mu_2$ and $\sigma_1 \neq \sigma_2$ and:

$$f_1(x) = \mathcal{N}(\mu_1, \sigma_1^2)$$

$$f_2(x) = \mathcal{N}(\mu_2, \sigma_2^2)$$

The Bhattacharyya coefficient is defined as:

$$BC(f_1, f_2) = \int \sqrt{f_1(x) f_2(x)}\mathrm{d}x, = \sqrt{\frac{2\sigma_1\sigma_2}{(\sigma_1^2 + \sigma_2^2)}} \exp\left\{\frac{-(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)}\right\}.$$

As a result, the Bhattacharyya distance D_B is:

$$\begin{aligned} D_B(f_1(x), f_2(x)) &= -\ln\left(BC(f_1(x), f_2(x))\right) = \\ -\ln&\left(\sqrt{\frac{2\sigma_1\sigma_2}{(\sigma_1^2 + \sigma_2^2)}} \exp\left\{\frac{-(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)}\right\}\right) = \\ &-\frac{1}{2}\ln\left(\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}\right) + \frac{1}{4}\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} = \\ &\frac{1}{2}\ln\left(\frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1\sigma_2}\right) + \frac{1}{4}\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}. \end{aligned}$$

For multivariate normal distributions, we will use covariance instead of varicance, $f_i = N(\mu_i, \Sigma_i)$ the Bhattacharyya distance will be:

$$D_B = \frac{1}{2}\ln\left(\frac{\det\Sigma}{\sqrt{\det\Sigma_1 \det\Sigma_2}}\right) + \frac{1}{8}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2),$$

where μ_i and Σ_i are the means and covariance of the distributions, and

$$\Sigma = \frac{\Sigma_1 + \Sigma_2}{2}.$$

As seen from these equations the Bhattacharyya distance is a generalization of the Mahalanobis distance. When the variances of the two distributions are the same the first term of the distance is zero as this term depends solely on the variances of the distributions, and the distance will be

$$D_B = \frac{1}{8}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2),$$

that is identical to the Mahalanobis distance between two means μ_1, μ_2.

On the other hand, if the variances are different and the means are equal (as shown in Fig. 1) the Mahalanobis distance will be zero, where the Bhattacharyya distance which takes into account the differences between the variances.

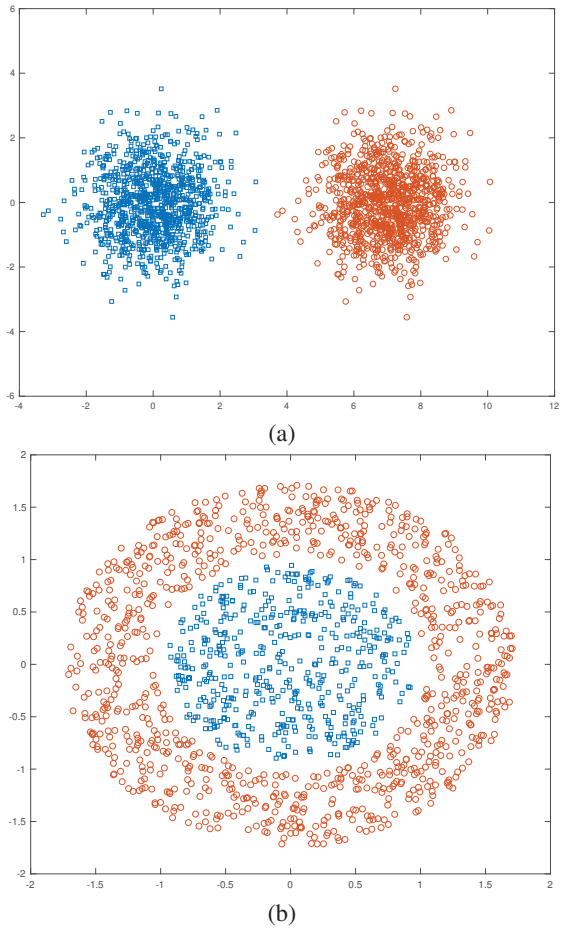


Fig. 1 Bhattacharyya distance for two special cases: (a) The variances are the same ($\sigma_1 = \sigma_2$), means are different ($\mu_1 \neq \mu_2$). (b) The means are equal ($\mu_1 = \mu_2$), variances are different ($\sigma_1 \neq \sigma_2$)

IV. Bhattacharyya Distance Based Disatnce Function Over Incomplete Datasets

We now turn to define the proposed distance metric [1] that we used over the incomplete diabetes and breast cancer datasets. Let A be a set of points (i.e., each point represent a patient profile), where each coordinate describes one parameter from the patient profile. Given a measured value x_i for the ith coordinate i c_i, the conditional probability for c_i will be $P(c_i|x_i) \sim \mathcal{N}(x_i, \sigma_i^2)$, where x_i is the mean and σ_i^2 is the variance of the sensor/risk factor which measured the coordinate c_i. When on the other hand the value of x_i is missing then the probability distribution for c_i might be given in advance or can be computed according to the known values for this coordinate from the data (i.e., $P(c_i) \sim \chi_i$), where χ is the distribution. In our derivation when the distribution is unknown we estimate it using the kernel density estimation method (KDE) from the measured values.

Note that since each specific coordinate is measured by the same sensor and under the same conditions, each coordinate has a specific variance σ_i^2. Our method can be generalized to deal with coordinates whose measurements are dependant, but for simplicity we assume that the coordinates measurements are independent. Under these assumptions we will treat each coordinate separately.

Given two sample points X and Y, the goal is to compute the distance between them. Let x_i and y_i be the ith coordinate values from points X, Y respectively. There are three possible cases for the values of x_i and y_i: (1) Both values are given. (2) One value is missing. (3) Both values are missing.

1) Two Values Are Known: When the values of x_i and y_i are given the distance between them will be defined as:

$$D_B(x_i, y_i) = DB(N(x_i, \sigma_{i1}^2), N(y_i, \sigma_{i2}^2)) = \frac{1}{2}\ln\left(\frac{\sigma_{i_1}^2 + \sigma_{i_2}^2}{2\sigma_{i_1}\sigma_{i_2}}\right) + \frac{1}{4}\frac{(x_i - y_i)^2}{\sigma_{i_1}^2 + \sigma_{i_2}^2}.$$

Since x_i and y_i were measured by the same sensor $\sigma_{i_1} = \sigma_{i_2} = \sigma_i$ and thus

$$D_B(x_i, y_i) = \frac{1}{8}\frac{(x_i - y_i)^2}{\sigma_i^2}. \quad (1)$$

As mentioned above, this is the Mahalanobis distance which is the standard distance measurement between two points. In this case, the runtime complexity is $O(1)$.

2) One Value Is Missing: Suppose that x_i is missing and the value y_i is given. Since the value of x_i is unknown, we can not compute its Bhattacharyya distance. Instead we model the distance as a random selection of a point from the distribution of its coordinate χ_i and compute its distance. The mean of this computation is our distance. We will estimate this value as follows: We divide the range of c_i $[\min(c_i), \max(c_i)]$ into $l-1$ equal intervals $(m_1, \ldots, m_l)$ as illustrated in Fig. 2.

For each value m_j we can estimate its probability density $p(m_j)$ using the KDE. The probability for the jth interval Δj is:

$$P(\Delta j) = p(m_j) \cdot \frac{\max(c_i) - \min(c_i)}{l-1}.$$

As a result, we approximate the Mean Bhattacharyya distance (MD_B) between y_i and the distribution as:

$$MD_B(\chi_i, y_i) = \sum_{j=1}^{l-1} P(\Delta_j) D_B(\mathcal{N}(m_j, \sigma_1), \mathcal{N}(y_i, \sigma_1)).$$

This metric measures the distance between y_i and each suggested value of x_i and takes into account the probability for this value according to the evaluated probability distribution.

This is in contrast to the **Most Common Attribute Value** method. There the value of the attribute that occurs most often is selected to be the value for all the unknown values of the attribute and imply that the probability of the most common attribute value is 1 and 0 for all other possible values. Furthermore our distance is different from the **Mean Attribute Value method**, where the mean of a specific attribute is selected to replace the unknown values of the attribute because it does not take into account the dispersion of the values in the distribution. Thus for example two distributions with the same mean and different variances (as can be seen in Fig. 3) will get the same distance whereas in our method the distance increases as a function of the variance.

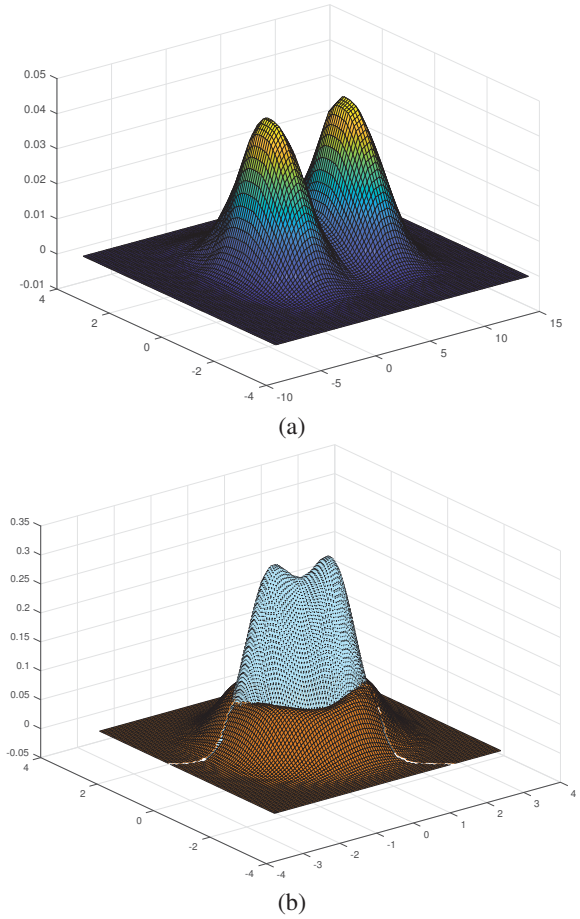


Fig. 2 An example for the normal kernel density estimation results for coordinate c_i. m_j denotes the selected points and $p(m_j)$ denotes the probability density for m_j

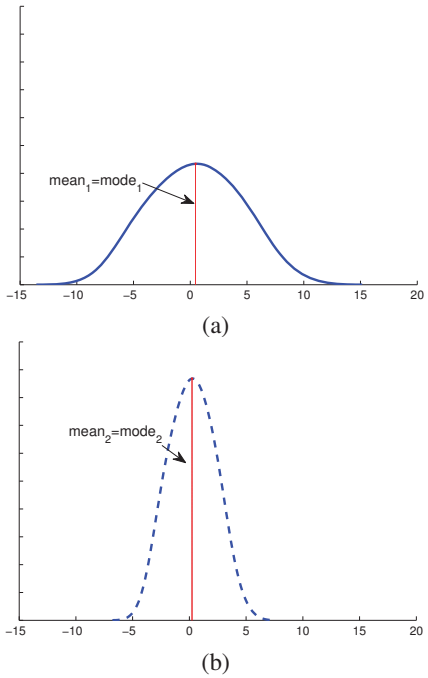


Fig. 3 (a) and (b) show two distributions with the same mean and different variances. The distance computed for these two distributions is different

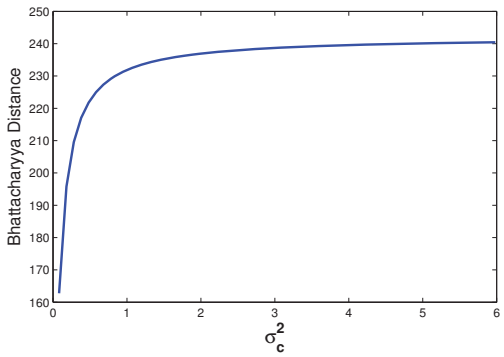


Fig. 4 The distance between a measured point and an unknown value with different values of the variance σ_c^2 of the distribution χ_i

Fig. 4 illustrates the dependance of our distance on the variance of distribution χ_i. When the variance is close to the measurement variance σ_i^2 the distance will converge to the value achieved for a measured value. As the variance increases the distance increases until it converges to the distance achieved for the uniform distribution.

In this case (i.e., one value is missing), the runtime of our method is $O(l)$, since according to this metric the algorithm has to compute $l-1$ Bhattacharyya distances. On the other hand as l increases so does the accuracy of the distance estimation. There fore, there is a trade off between the accuracy of the estimate and the the complexity of the algorithm. From our experiments we did not find a significant change in the performance of the classification algorithms as a function of l.

3) The Two Values Are Missing: In this case in order to estimate the Mean Bhattacharyya Distance we have to randomly select values for both x_i and y_i. Both of these values are selected from distribution χ_i. In order to compute the mean the following double sum has to be computed.

$$MD_B = \sum_{q=1}^{l-1}\sum_{j=1}^{l-1} P(\Delta_{1q})P(\Delta_{2j})DB(\mathcal{N}(m_{1q},\sigma_i),\mathcal{N}(m_{2j},\sigma_i)).$$

Consider again the examples in Fig. 3. The MD_B of the first distribution with the larger variance will naturally be larger than the MD_B of the second distribution with the smaller variance. Fig. 5 shows the dependance of MD_B on the variance σ_c^2 of the distribution χ_i. As the distribution is more dispersed, the value of the MD_B increases. In this example the distributions χ_i were Gaussian but the relationship is general.

As in this case no value has to be known in order to compute the MD_B the distance between two missing values from a specific coordinate will be fixed, and has to be computed only

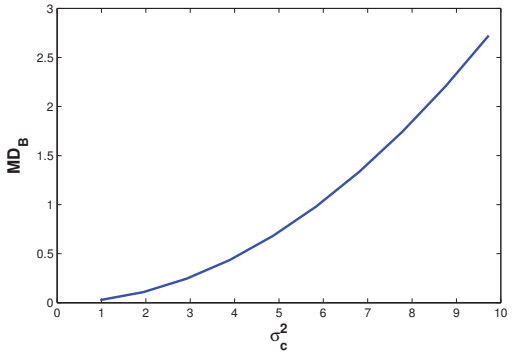


Fig. 5 The value of MD_B as a function of the variance σ_c^2 of the distribution χ_i

once. It therefore does not have any effect on the runtime of the algorithm.

V. Experiments on Numerical Datasets

In order to measure the ability of the new distance function to reflect the actual similarity or dissimilarity between instances with missing values we compare the performance of the kNN ($k = 1$) classifier on complete data (i.e., without missing values) to the performance of the kNN classifier using our distance (KNN-BH), the kNN-MC (i.e., Most Common attribute value), the kNN-MA (i.e., the Mean value of each Attribute), and the kNN-MI(Mean Imputation) that replaces a data point with missing values with the mean of all the instances in the data, on the same datasets with missing values.

We ran our experiments on health standard numerical datasets for diabetes and breast cancer diseases from the Machine Learning Repository (UCI) [5]. The first dataset is the Pima Indians diabestes datasets. The owner of this dataset is the National Institute of Diabetes and Digestive and Kidney Diseases. In particular, all patients here are females at least 21 years old of Pima Indian heritage. This data contains 762 patients and 8 attributes for each woman as follow:

1) Number of times pregnant
2) Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3) Diastolic blood pressure (mm Hg)
4) Triceps skin fold thickness (mm)
5) 2-Hour serum insulin (mu U/ml)
6) Body mass index (weight in $kg/(heightinm)^2$)
7) Diabetes pedigree function
8) Age (years)

The second is the Wisconsin Diagnosis Breast Cancer dataset which contains 683 patients and 8 attributes as follow:

1) Clump Thickness
2) Uniformity of Cell Size
3) Uniformity of Cell Shape
4) Marginal Adhesion
5) Single Epithelial Cell Size
6) Bare Nuclei
7) Bland Chromatin
8) Normal Nucleoli
9) Mitoses

Both of these data sets are two-classe classification problem. The characteristics of all the datasets can be seen in Table I. Those datasets were labeled, but this knowledge was used only to evaluate the accuracy of the resulting classifier. In all experiments these datasets are assumed to be unlabeled.

TABLE I
Dataset Properties

Dataset	Dataset size	Classes
Pima Indians	762×8	2
Breast Cancer	683×8	2

In the first stage of the experiments, from each dataset a set of size 10%-50% of the dataset is randomly drawn to be samples with missing values, where at least one coordinate from each instance was selected randomly to be the missing value. After that, from each dataset a set of 10% of the dataset was drawn randomly to be the training dataset (i.e., labeled) and the rest is the testing dataset. (Note that the training dataset may contains instances with missing values.) Then the accuracy was evaluated for each set of missing values by the ability of the kNN classifier to label the data. The results are averaged over 10 different runs on each dataset. A resulting curve was constructed for each dataset to evaluate how well the algorithm performed.

A. Results

As can be seen from Fig. 6, the kNN-BH was superior and outperforms the other algorithms. The learning curves are constructed by computing the ratio of correctly classified instances to the whole unlabeled data.

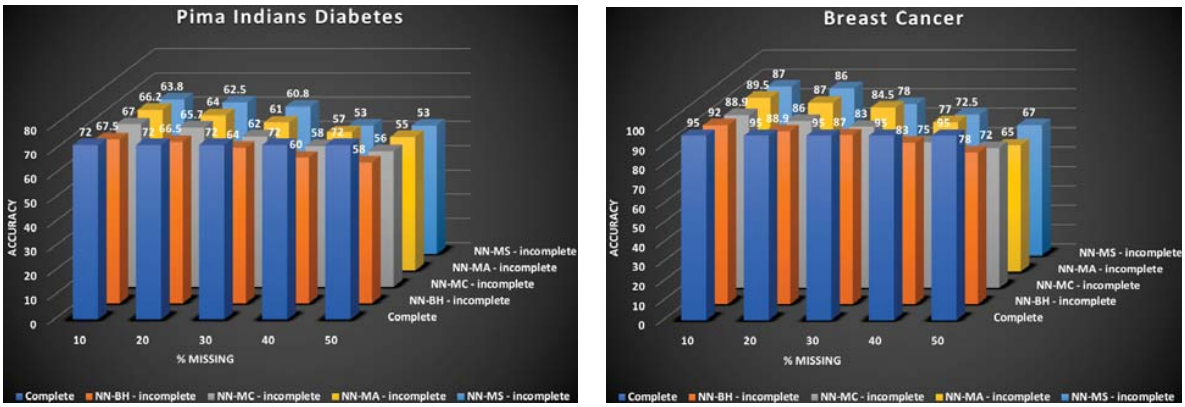
The main goal here is to compare our method to the exists methods that deal with the missing data problem. As can be seen in the results curves (in Fig. 6), the kNN-BH obviously outperforms the other methods. Moreover, according to the results curves the performances of the kNN-MC was better than the performance of the kNN-MA over the Pima Indians dataset, while on the Breast Cancer datasets the performance of the kNN-MA was better. In both datasets the performance of the kNN-MS was poorly.

This improvement in kNN-BH performanct proofs the ability of the derived method that based on the Bhattacharyya distance to better measure the actual similarity/dissimilairty between the different objects with missing values.

VI. Conclusions

Many real-world datasets suffer from the problem of missing values. Several methods have been proposed to measure the similarity between objects with missing values. In this research, we derived a new distance function based on data attributes distribution using the Bhattacharyya distance and used is for incomplete datasets. The developed distance distinguishes between two cases: (a) complete points and (2) incomplete points.

To measure its ability to measure the similarity between different objects we integrated it within the frame work of the kNN classifier framework (we use the one nearest neighbor



(a) (b)

Fig. 6 Results of 1NN without missing values, 1NN-BH, 1NN-MC, 1NN-MS and 1NN-MA algorithms over six numerical datasets with missing values

classifier). We use standard benchmark data from the UCI repository for both diabetes and breast cancer diseases. From our experiment we conclude that our distance is a more appropriate function to measure the similarity between objects with missing value especially when the percent of the missing values is becomes large. This is because when the missing data is small, the missing value does not influence the similarity value significantly.

This distnace is general and can be used as part of many machine learning algorithm that used the distance between data points.

Acknowledgment

This research was supported by Wikaya ltd company.

References

[1] L. Abedallah and I. Shimshoni. A distance function for data with missing values and its application. *Proc. of the 13th Int. Conf. on Data Mining and Knowledge Engineering*, 2013.

[2] G. Batista and M.C. Monard. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5-6):519–533, 2003.

[3] Krzysztof J Cios and Lukasz A Kurgan. Trends in data mining and knowledge discovery. *Advanced techniques in knowledge discovery and data mining*, pages 1–26, 2005.

[4] A Rogier T Donders, Geert JMG van der Heijden, Theo Stijnen, and Karel GM Moons. Review: a gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091, 2006.

[5] A. Frank and A. Asuncion. UCI machine learning repository at http://archive.ics.uci.edu/ml. visited (2013), 2010.

[6] Jerzy Grzymala-Busse and Ming Hu. A comparison of several approaches to missing attribute values in data mining. In *Proc. Rough Sets and Current Trends in Computing*, pages 378–385. Springer, 2001.

[7] Joseph G Ibrahim, Ming-Hui Chen, Stuart R Lipsitz, and Amy H Herring. Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346, 2005.

[8] Roderick JA Little. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287–296, 1988.

[9] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[10] Matteo Magnani. Techniques for dealing with missing data in knowledge discovery tasks. *Obtido http://magnanim.web.cs.unibo.it/index.html*, 15(01):2007, 2004.

[11] S. Zhang, Z. Qin, C.X. Ling, and S. Sheng. Missing is useful": missing values in cost-sensitive decision trees. *IEEE Trans. on KDE*, 17(12):1689–1693, 2005.

[12] Shichao Zhang. Shell-neighbor method and its application in missing data imputation. *Applied Intelligence*, 35(1):123–133, 2011.

Loai AbdAllah received his B.Sc. in Mathematics and Management Information Systems from the University of Haifa, his M.Sc. in Mathematics from the University of Haifa, where he is currently working toward the Ph.D. degree in Mathematics in the University of Haifa. Loai was a member of the Departments of Mathematics and Computer Science at the College of Sakhnin from in October 2011. His current research interest is in data mining.

Mahmoud Kaiyal a family doctor graduated from The Hebrew University Hadassah Medical School (MD). After 10 years of practicing family medicine he turned to digital health, founded WebTeb, a medical and health portal in Arabic providing digital health solutions for the all healthcare stakeholders in the region. In 2017 he cofounded WIKAYA, an AI-based platform for improving prevention to chronic diseases