CM-test: An Innovative Divergence Measurement and Its Application in Diabetes Gene Expression Data Analysis

Lily R. Liang, Shiyong Lu, *Member, IEEE*, Yi Lu, Puneet Dhawan and Deepak Kumar

***Abstract*— One important problem in data analysis is to effectively measure the divergence of two sets of values of a feature, each from a group of samples with a particular condition. Such a measurement is the foundation for identifying critical features that contribute to the difference between the two conditions. The two traditional methods t-test and Wilcoxon rank sum test measure this divergence indirectly, using the difference of the means of the two groups and the sum of the ranks from one of the groups, respectively. In this paper, we propose an innovative approach based on fuzzy set theory, the Cluster Misclassification test (CM-test), to quantify the divergence directly and robustly.**

To validate our approach, we conducted experiments on both synthetic and real diabetes gene expression datasets. On the synthetic datasets, we observed that CM-test effectively quantifies the divergence of two sets. On the real diabetes dataset, we observed that in the top ten genes identified by CM-test, eight of them have been confirmed to be associated with diabetes in the literature. We suggest the remaining two genes, M95610 and M88461, as two potential diabetic genes for further biological investigation. Therefore, we recommend that CM-test be another effective method for measuring the divergence of two sets, complementing t-test and Wilcoxon rank sum test in practice.

***Index Terms*— CM-test, Divergence Measurement, Fuzzy Sets, Gene Expression Analysis**

I. Introduction

One important problem in data analysis is to effectively measure the divergence of two sets of values of a feature, each from a group of samples with a particular condition. Such a measurement is the foundation for identifying critical features that contribute to the difference between the two conditions.

A motivating example is shown in Table I, which records the microarray gene expression values of five genes for two groups of people: five insulin-sensitive humans and five insulin-resistant humans. In order to identify the genes that are associated with diabetes, one needs to determine for each gene whether or not the two sets of expression values are significantly different from each other. One popular method is t-test [19], which uses the difference of the means of the two sets to measure the divergence. In Table I, the first three genes are identified by t-test as significant genes (with p-value ≤ 0.05). However, one limitation of t-test is that it is only applicable when the two samples are normally distributed. Another limitation of t-test is that it is very sensitive to extreme values. As a result, t-test fails to recognize genes 4 and 5 as significant genes although their expression values under the two conditions are significantly different from each other. Another popular method is Wilcoxon rank sum test [19], which is based solely on the order in which the observations from the two samples fall. In Table I, the first four genes are identified by Wilcoxon rank sum test as significant genes (with p-value ≤ 0.05). In contrast to t-test, Wilcoxon rank sum test is applicable to samples with any distribution and is less sensitive to extreme values. However, the rank sum test is less sensitive to a location shift. As a result, Wilcoxon rank sum test fails to identify gene 5 as significant gene although the two sets for gene 5 are significantly different from each other.

In this paper, based on the fuzzy set theory [14], we propose an innovative approach that overcomes the above limitations of t-test and rank sum test. The basic idea is to consider the two sets of values as samples from two different fuzzy sets. We then examine the membership value of each element with respect to each of these two fuzzy sets. If an element belongs more to the other fuzzy set, then we say that the element is *misclassified*. By counting the number of misclassified elements and quantifying the degree of misclassification, we measure the divergence of the original two sets.

The main contributions of this paper are:

1) We propose an innovative approach based on the fuzzy set theory, the Cluster Misclassification test (CM-test), which quantifies the divergence of two sets directly.
2) We validate CM-test on synthetic datasets and show that it is effective and robust.
3) We apply CM-test to a real diabetes dataset and identified 10 significant genes, eight of which have been know to be associated with diabetes, with the remaining two genes suggested for further biological investigation.

The rest of the paper is organized as follows. Section II briefly reviews t-test and Wilcoxon rank sum test and their limitations. Section III presents our fuzzy-set-theory-based method, CM-test. Section IV provides our experimental results both on synthetic datasets and a real dataset of diabetes gene expression profiles. Finally, Section V concludes the paper and

This work was supported by the Agricultural Experiment Station at the University of the District of Columbia (Project No.: DC-0LIANG; Accession No.: 0203877)

Lily R. Liang is with the Department of Computer Science and Information Technology, University of the District of Columbia, (Email:lliang@udc.edu)

Shiyong Lu and Yi Lu are with the Department of Computer Science, Wayne State University, Detroit, MI 48202, (Email: {shiyong, luyi}@wayne.edu)

Puneet Dhawan is with the Department of Biochemistry and Molecular Biology, University of Medicine and Dentistry of New Jersey

Deepak Kumar is with the Department of Biology and Environmental Sciences, University of the District of Columbia

TABLE I

THE MICROARRAY GENE EXPRESSION VALUES FOR FIVE GENES FOR TWO GROUPS OF SUBJECTS UNDER TWO CONDITIONS

Gene ID	Condition 1					Condition 2					CM d-value	CM-test p-value	t-test p-value	rank sum p-value
1	750	559	649	685	636	310	359	135	97	178	1	0.005	0.00	0.008
2	391	379	268	323	380	774	506	416	468	449	1	0.005	0.029	0.008
3	234	272	275	201	231	82	150	132	**202**	146	0.910	0.008	0.003	0.015
4	233	216	193	**394**	327	436	980	363	424	416	0.905	0.017	0.071	0.015
5	598	424	695	451	141	**342**	260	266	229	234	0.904	0.018	0.077	0.152

points out some potential future work.

II. RELATED WORK

In the following, we briefly review the following: i) two most popular methods to measure the divergence of two sets of values, t-test [19] and Wilcoxon rank sum test [19]; ii) fuzzy set theory [24] and fuzzy clustering validity indices [11].

A. Statistical methods: t-test and Wilcoxon rank sum test

The statistical method t-test assesses whether the means of two groups are statistically different from each other. Given two sets S_1 and S_2, we assume that S_1 is a random sample of size n_1 from an $N(\mu_1, \sigma_1^2)$ distribution, and S_2 is a random sample of size n_2 from an $N(\mu_2, \sigma_2^2)$ distribution, and $\sigma_1^2 \neq \sigma_2^2$. We wish to test the hypothesis $H_0 : \mu_1 = \mu_2$ versus $\mu_1 \neq \mu_2$. The t statistic is calculated as

$$t(S_1, S_2) = \frac{\mu_{S_1} - \mu_{S_2}}{\sqrt{\frac{\sigma_{S_1}^2}{n_1} + \frac{\sigma_{S_2}^2}{n_2}}} \quad (1)$$

where μ_S and σ_S are the sample mean and standard deviation of S, respectively.

The exact distribution of t under H_0 is difficult to derive. The Satterthwaite approximation computes the approximate degress of freedom d' as follows.

$$d' = \frac{(\sigma_{S_1}^2/n_1 + \sigma_{S_2}^2/n_2)^2}{(\sigma_{S_1}^2/n_1)^2/(n_1-1) + (\sigma_{S_2}^2/n_2)^2/(n_2-1)} \quad (2)$$

The method then rounds d' down to the nearest integer d''. If $t > t_{d'',1-\alpha/2}$ or $t < -t_{d'',1-\alpha/2}$ then reject H_0, otherwise accept H_0.

The primary limitation of t-test is that it requires the two samples to be normally distributed. Another limitation of t-test is that it is very sensitive to extreme values.

Another popular statistical method is Wilcoxon rank sum test, which can be used to test the null hypothesis that two sets S_1 and S_2 have the same distribution. We first merge the data from these two sets and rank the values from the lowest to the highest with all sequences of ties being assigned an average rank. The Wilcoxon test statistic W is the sum of the ranks from set S_1. Assuming that the two sets have the same continuous distribution (and no ties occur), then W has a mean and standard deviation given by

$$\mu = \frac{m * (m + n + 1)}{2} \quad (3)$$

$$\sigma = \sqrt{\frac{m * n * (m + n + 1)}{12}}, \quad (4)$$

where $m =\mid S_1 \mid$ and $n =\mid S_2 \mid$.

We test the null hypothesis H_o: no difference in distributions. A one-sided alternative is H_a: S_1 yields lower measurements. We use this alternative if we expect or see that W is unusually lower than its expected value μ. In this case, the p-value is given by a normal approximation. We let $N \sim N(\mu, \sigma)$ and compute the left-tail $P(N \leq W)$ (using continuity correction if W is an integer).

If we expect or see that W is much higher than its expected value, then we should use the alternative H_a: first S_1 yields higher measurements. In this case, the p-value is given by the right-tail $P(N \geq W)$. If the two sums of ranks from each set are close, then we could use a two-sided alternative H_a: there is a difference in distributions. In this case, the p-value is given by twice the smallest tail value $2P(N \leq W)$ if $W < \mu$, or $2P(N \geq W)$ if $W > \mu$.

In contrast to t-test, rank sum test does not require the two samples to be normally distributed and is less sensitive to extreme values. However, it is less sensitive to a location shift. This might be a disadvantage for some applications.

B. Fuzzy set theory and fuzzy clustering validity indices

Fuzzy set theory is an extension of the conventional (crisp) set theory. It uses the concept of partial truth to model vagueness and ambiguity. It was introduced by Prof. Lotfi A. Zadeh at UC Berkeley in 1965. In this theory, a fuzzy set is a class of objects with a continuum of grades of membership. These grades of membership range between zero and one and are assigned by a membership function which characterizes the fuzzy set. Fuzzy sets do not have to be disjoint and when they are not, their fuzzy membership functions overlap. So one object can belong to two different fuzzy sets with different grades of membership.

Fuzzy set theory has been applied to fuzzy clustering [12]. A clustering procedure partitions data into clusters such that similar data objects belong to the same cluster and dissimilar data objects to different clusters. However, In real applications there is very often no sharp boundary between clusters so that fuzzy clustering is often better suited for the data. Membership degrees between zero and one are used in fuzzy clustering instead of crisp assigment of the data to clusters. Popular fuzzy clustering algorithms include fuzzy c-means [6], [1], Gustafson-Kessel algorithm [10], Gath-Geva algorithm (GG) [9], fuzzy c-varieties algorithm, adaptive fuzzy c-varieties algorithm, fuzzy c-shells algorithm, fuzzy c-spherical shells algorithm, fuzzy c-rings algorithm, fuzzy c-quadric shells algorithm, fuzzy c-rectangular shells algorithm; see [12] for more details of these algorithms.

One of the most important issues in cluster analysis is the evaluation of clustering results to find the partitioning that best fits the underlying data. A clustering validity index provides a goodness-of-clustering value. Some general indices for fuzzy clusters are separation indices D1 and D2 [5], the partition coefficient (PC) [1], the classification entropy (CE) [2] and Xie-Beni index [22]. These indices either involves only membership values or involves the clustered data as well. Like other non-fuzzy clustering indices, they are usually used for the evaluation of the clustering results at the end of a clustering process, when each object has already been assigned to the fuzzy cluster for which the object has the highest grade of membership.

For the sample microarray data in Table I, the five values of each condition can be considered as a fuzzy set, thus, the fuzzy set theory applies. However, these two sets are not the results of clustering, instead, they are the result of labeling of the values at the time they were read. Thus, there is a great chance that a value in its labeled fuzzy set belongs more to the other fuzzy set. The CM-test we propose in this paper takes advantage of this particular property.

III. Methodology

In this section, based on the fuzzy set theory [14], we present our innovative approach, the Cluster Misclassification test (CM-test), to quantify the divergence of two sets of values directly and robustly.

Let S_1 and S_2 be two sets of values of a particular feature for two groups of samples under two different conditions. The basic idea is to consider the two sets of values as samples from two different fuzzy sets. We examine the membership value of each element with respect to each of these two fuzzy sets. If an element belongs more to the other fuzzy set, then we say that the element is *misclassified*. By counting the number of misclassified elements and quantifying the degree of misclassification, we measure the divergence of the original two sets. In particular, we perform the following steps:

1) Compute the sample mean and standard deviation of S_1 and of S_2 respectively.
2) Characterize S_1 and S_2 as two fuzzy sets FS_1 and FS_2 whose fuzzy membership functions, $f_{FS_1}(x)$ and $f_{FS_2}(x)$, are defined with the sample means and standard deviations. The fuzzy membership function $f_{FS_i}(x)$ $(i = 1, 2)$ maps each value x to a fuzzy membership value that reflects the degree of x belonging to FS_i $(i = 1, 2)$.
3) Using the two fuzzy membership functions, $f_{FS_1}(x)$ and $f_{FS_2}(x)$, quantify the misclassification degree between these two sets.
4) Finally, define the cluster misclassification divergence degree (CM d-value) between the two sets based on the misclassification degree.

The details of each step is elaborated in the sequel.

A. Fuzzy Sets and Membership Functions

The sample mean μ_1 of S_1 is calculated as

$$\mu_1 = \frac{1}{n_1} \sum_{x_i \in S_1} x_i \tag{5}$$

where n_1 is the number of elements in S_1, and the sample standard deviation σ_1 of S_1 is calculated as

$$\sigma_1 = \sqrt{\frac{1}{n_1 - 1} \sum_{x_i \in S_1} (x_i - \mu_1)^2} \tag{6}$$

For gene 5 in Table I, we have $\mu_1 = 461.8$, $\sigma_1 = 210.59$, $\mu_2 = 266.2$, and $\sigma_2 = 45.29$. We then characterize set S_1 by a fuzzy set FS_1 whose fuzzy membership function is defined as

$$f_{FS_1}(x) = e^{-(x-\mu_1)^2/(2\sigma_1^2)} \tag{7}$$

The function $f_{FS_1}(x)$ maps each value x in S_1 to a fuzzy membership value to quantifies the degree that x belongs to FS_1. A value equal to the mean has a membership value of 1 and belongs to fuzzy set FS_1 to a full degree; a value that deviates from the mean has a smaller membership value and belongs to FS_1 to a smaller degree. The further the value deviates from the mean, the smaller the fuzzy membership value. Similarly, the fuzzy membership function for S_2 is defined as

$$f_{FS_2}(x) = e^{-(x-\mu_2)^2/(2\sigma_2^2)} \tag{8}$$

where μ_2 and σ_2 are the mean and standard deviation of S_2 respectively.

For gene 5 in Table I, we have $f_{FS_1}(x) = e^{-(x-461.8)^2/88696.3}$ and $f_{FS_2}(x) = e^{-(x-266.2)^2/4102.4}$. With these two fuzzy membership functions, the fuzzy membership values for each element with respect to the two sets can be calculated. For example, $f_{FS_1}(598) = 0.81$ and $f_{FS_2}(598) = 2.2E^{-12}$.

B. Our Proposed Method CM-test

Since the fuzzy membership functions can overlap, one element can belong to more than one fuzzy set with a respective degree for each. For an element in S_1, we measure the degree that it belongs to FS_1 by applying its value to f_{FS_1}. Similarly we can apply its value to f_{FS_2} to measure the degree that it belongs to FS_2. We say an element in S_1 is *misclassified* if it belongs more to FS_2 in the sense of fuzzy membership value, and vice versa. The idea of CM-test is to aggregate the number of misclassified elements as well as the degree of misclassification of elements in both S_1 and S_2. First, we define the notion of *element misclassification degree* as follows.

Definition 3.1 (Element misclassification degree): Given two sets S_1 and S_2 and their corresponding fuzzy set FS_1 and FS_2, the element misclassification degree of an element e in S_1 with respect to FS_2 is defined as

$$m(e, FS_2) = \begin{cases} f_{FS_2}(e) - f_{FS_1}(e) & : \ if\ f_{FS_2}(e) > f_{FS_1}(e) \\ 0 & : \ otherwise \end{cases} \tag{9}$$

□

For gene 5 in Table I, since $f_{FS_1}(598) = 0.81$ and $f_{FS_2}(598) = 2.2E^{-12}$, we have $f_{FS_1}(598) > f_{FS_2}(598)$, i.e., 598 belongs more to its own cluster than to the other cluster. Therefore, $m(598, FS_2) = 0$. On the other hand, since $f_{FS_2}(342) = 0.246$, $f_{FS_1}(342) = 0.851$, we have $f_{FS_1}(342) > f_{FS_2}(342)$, i.e., 342 belongs more to the other cluster than to its own cluster. Hence, it is misclassified, and we have $m(342, FS_1) = f_{FS_1}(342) - f_{FS_2}(342) = 0.605$.

We denote the misclassified elements in S_1 with respect to FS_2 as $M_{FS_2}(S_1) = \{e \mid e \in S_1 \wedge m(e, FS_2) > 0\}$. Similarly, we denote the misclassified elements in S_2 with respect to FS_1 as $M_{FS_1}(S_2) = \{f \mid f \in S_2 \wedge m(f, FS_1) > 0\}$. Finally, we denote the number of misclassified elements in S_1 and S_2 with respect to each other as $\#M(S_1, S_2) = \mid M_{FS_2}(S_1) + M_{FS_1}(S_2) \mid$. For gene 5 in Table I, the only misclassified element is 342, thus we have $M_{FS_2}(S_1) = \{\}$, $M_{FS_1}(S_2) = \{342\}$, and $\#M(S_1, S_2) = 1$. All the misclassified elements are bolded in Table I.

We then define the convergence degree of two S_1 and S_2 as a linear interpolation of two terms: the number of misclassified elements and the mutual misclassification degrees.

Definition 3.2 (CM c-value): Given two sets S_1 and S_2, the cluster misclassification convergence degree (CM c-value) between S_1 and S_2 is defined as

$$c(S_1, S_2) = \alpha * T_1 + (1 - \alpha) * T_2 \tag{10}$$

where

$$T_1 = \frac{\#M(S_1, S_2)}{\mid S_1 \mid + \mid S_2 \mid}, \tag{11}$$

$$T_2 = \frac{\sum_{e \in S_1} m(e, S_2) + \sum_{f \in S_2} m(f, S_1)}{\mid S_1 \mid + \mid S_2 \mid}, \tag{12}$$

and α is an algorithm parameter whose value ranges from 0 to 1. □

In our implementation, we choose $\alpha = 0.9$ since we like to rank a gene with a greater number of misclassified elements before those that have a smaller number of misclassified elements, although other α values can be used for different applications. For gene 5 in Table I, only one element 342 is misclassified with element misclassification degree of 0.605. We have $T_1 = 0.1$ and $T_2 = 0.0605$. Thus the CM c-value is $c(S_1, S_2) = 0.9 * 0.1 + 0.1 * 0.0605 = 0.09605$. Finally, we define the cluster misclassification divergence value (CM d-value) as follows.

Definition 3.3 (CM d-value): Given two sets S_1 and S_2, the cluster misclassification divergence degree (CM d-value) between S_1 and S_2 is defined as

$$d(S_1, S_2) = 1 - c(S_1, S_2) \tag{13}$$

□

For gene 5 in Table I, $c(S_1, S_2) = 0.09605$, thus the CM d-value is $1 - c(S_1, S_2) = 0.90395$. We have calculated all the p-values for the five genes in Table I for the three methods. One interesting observation is that, while t-test identifies gene 3 is more significant than gene 2, both CM-test and Wilcoxon rank sum test rank gene 2 higher than gene 3. This is more reasonable as in gene 2, none of the elements is misclassified, and thus the two clusters are fully separated, while in gene 3, one element is misclassified. Another interesting observation is that, while both t-test and Wilcoxon rank sum test fail to recognize gene 5 as a significant gene since their p-values are greater than 0.05, our CM-test identifies gene 5 as a significant gene with a p-value of 0.018. This is more reasonable since only one element is misclassified in this gene expression set pair. The reason of the failure of t-test and Wilcoxon rank sum test is due to their sensitivity to the extreme value 141 in the first set of the gene.

C. Interpreting CM d-value with Empirical p-value

Given a calculated CM d-value D for two sets S_1 and S_2, to interpret D in terms of "significantly divergent" or not, we need to know the cutoff value δ of D, so that when $D \geq \delta$, the two sets are interpreted as significantly divergent. To define this cutoff value, we resort to the probability theory. In the context of CM-test, we like to test the following null hypothesis H_o: S_1 and S_2 originate from the same distribution. Then the p-value is defined as the probability $P(d(S_1, S_2) \geq D \mid S_1$ and S_2 were randomly sampled from the same distribution$\}$. The p-value answers the following question: if S_1 and S_2 were randomly drawn from the same distribution, what is the probability that the CM d-value of these two sets would be equal or greater than the observed value D? As a convention of statistical analysis, if $p - value \leq 0.05$, then it is a strong evidence to reject the null hypothesis, and accepts that the two sets are significantly divergent, while the p-value reflects the significance – the smaller the p-value, the greater the significance. It is not easy to calculate the real p-value for CM d-value due to the following reasons: (1) it is still not clear whether or not CM d-value has a standard asymptotic distribution; (2) even if a standard asymptotic distribution does exist, it may not be reliable in our small sample size; and (3) the calculation of the exact sampling distribution through exhaustive enumeration of all possible samples may be too computationally intensive to be feasible. It has been very common to use Monte Carlo procedures to calculate the empirical p-value which approximates the exact p-value without relying on asymptotic distributional theory or on exhaustive enumeration. Davison and Hinkley [4] present the formula for obtaining an empirical p-value as $(n+1)/(N+1)$, where N is the number of samples in the data set, and n is the number of those samples which produce the statistical value greater than or equal to the specified value. In the context of CM-test, we perform the following steps to calculate the p-value of two sets S_1 and S_2 with their CM d-value D:

1) Estimate the distribution that S_1 and S_2 are drawn from as a normal distribution $N(\mu, \sigma)$, where μ and σ are estimated using the sample mean and standard deviation of $S_1 \cup S_2$.
2) Randomly draw N pairs of sets from $N(\mu, \sigma)$, then calculate the CM d-value for each pair. Let n be the number of pairs whose CM d-values are equal or greater than D.

3) Calculate the empirical p-value as $(n+1)/(N+1)$.

The cutoff CM d-value we obtain in this way is introduced in the next section.

IV. Experimental Results and Discussion

To validate our approach, first, we conducted CM-test on a synthetic dataset to study the relationship between CM d-value and its empirical p-value. Second, we conducted CM-test on another synthetic dataset to study the relationship between CM d-value and the mean difference of distributions. Finally we conducted CM-test on a real microarray dataset of diabetes gene expressions to identify genes that are related to diabetes and insulin metabolism

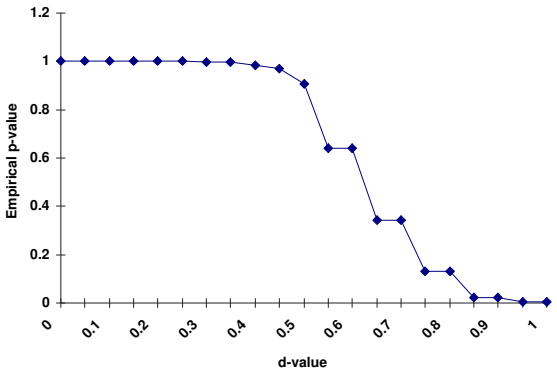

Fig. 1. The relationship between CM d-value and empirical p-value

A. Relationship between CM d-value and empirical p-value

Suppose two sets S_1 and S_2 are drawn from the same normal distribution, then what is the probability that they have a CM d-value equal to or greater than a particular D? And as D increases, will this probability decrease?

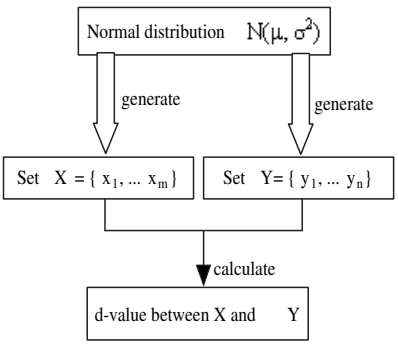

Fig. 2. Random generation of one d-value from one normal distribution

To answer the above questions, we studied the relationship between CM d-value and empirical p-value as follows:

1) We generated $N = 10000$ pairs of sets of values, with each set containing 5 values. As shown in Figure 2, each value in both sets is randomly generated from the standard normal distribution $N(0, 1)$.
2) We calculated the CM d-value for each pair of sets.
3) For each pair of sets S_1 and S_2 with CM d-value D, we calculated its empirical p-value as $n/10000$ where n is the number of pairs in these 10000 pairs that have a CM d-value equal to or greater than D.
4) Finally, we drew the relationship between CM d-value and empirical p-value in Figure 1.

From Figure 1, we can see that as CM d-value increases, the p-value decreases. In particular, when $D \geq 0.8136$, $p - value \leq 0.05$. Therefore, given two sets S_1 and S_2 drawn from the same normal unit distribution, the chance that the pair has a CM d-value equal to or greater than 0.8136 is very low. On the other hand, if we observe that two sets have a CM d-value equal to or greater than 0.8136, then there is a strong evidence that these two sets are drawn from two different distributions. Therefore, they should considered as significantly divergent.

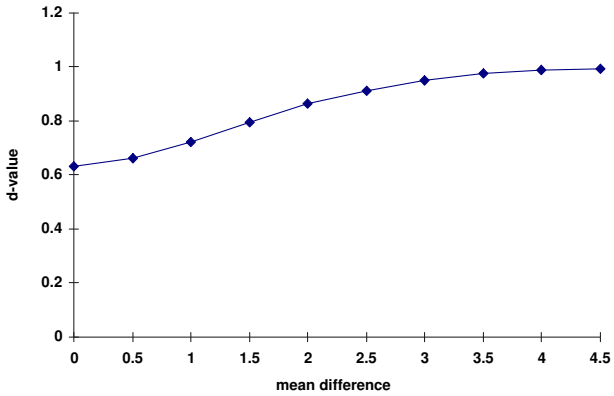

Fig. 3. Relationship between the mean difference of distributions and CM d-value

B. Relationship between the mean difference of distributions and CM d-value

Suppose two sets S_1 and S_2 are drawn from two different distributions, then a good divergence measurement should satisfy the following property: the less overlap between these two distributions, the greater the CM d-value.

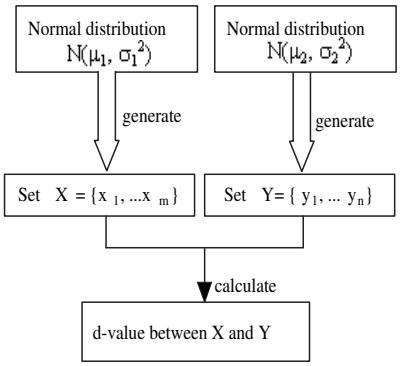

Fig. 4. Random generation of one d-value from two normal distributions

TABLE II

TEN BEST-RANKED AND TEN WORST-RANKED GENES IDENTIFIED BY CM-TEST

Probe Set	Gene Description	CM d-value	Empirical p-value	t-test p-value	rank sum p-value
U45973	Human phosphatidylinositol (4,5) bisphosphate 5-phosphatase homolog	1	0.005	0.00	0.01
M60858	Human nucleolin gene	1	0.005	0.00	0.01
M95610	Human alpha 2 type IX collagen (COL9A2) mRNA	1	0.005	0.01	0.01
L07648	Human MXI1 mRNA	1	0.005	0.01	0.01
L07033	Human hydroxymethylglutaryl-CoA lyase mRNA	1	0.005	0.01	0.01
X53586	Human mRNA for integrin alpha 6	1	0.005	0.01	0.01
X81003	Homo sapiens HCG V mRNA	1	0.005	0.01	0.01
L27559	Human insulin-like growth factor binding protein 5 (IGFBP5) gene, partial exon 4	1	0.005	0.03	0.01
M88461	Human neuropeptide Y peptide YY receptor mRNA	1	0.005	0.03	0.01
U65785	Human 150 kDa oxygen-regulated protein ORP150 mRNA	0.91	0.007	0.02	0.01
X55954	ribosomal protein L23	0.361	0.996	0.572	0.551
X57766	matrix metallopeptidase 11 (stromelysin 3)	0.361	0.996	0.693	0.842
U26173	nuclear factor, interleukin 3 regulated	0.361	0.996	0.483	0.970
D80008	DNA replication complex GINS protein PSF1	0.360	0.997	0.447	0.555
L35475	olfactory receptor, family 2, subfamily H, member 2	0.360	0.998	0.524	0.693
M69066	moesin	0.359	0.998	0.403	0.839
M97935	signal transducer and activator of transcription 1, 91kDa	0.359	0.998	0.373	0.839
L20348	oncomodulin	0.359	0.998	0.405	0.544
X61072	T cell receptor alpha joining 31	0.273	0.999	0.750	0.837
D50063	proteasome (prosome, macropain) 26S subunit, non-ATPase, 7 (Mov34 homolog)	0.268	1	0.543	0.421

We validated that our CM-test has this property as follows:

1) Let $N(0,1)$ and $N(x,1)$ be two normal distributions, where x is the mean difference between these two distributions. In this experiment, we consider x = 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, respectively.
2) As shown in Figure 4, we generated $N = 1000$ pairs of sets of values, with the first set containing 5 values that are randomly generated from $N(0,1)$, and the second set containing 5 values that are randomly generated from $N(x,1)$.
3) We calculated the CM d-value for each pair. Let the average of these 1000 CM d-values be d. We then plotted (x,d) in Figure 3.
4) We repeated step 2 and 3 for different x. Finally, the curve was drawn in Figure 3.

Figure 3 confirmed the desirable property of CM-test: the larger the mean difference between the two distributions, the greater the CM d-value.

C. Applying CM-test in Diabetes Gene Expression Analysis

A dataset of Microarray gene expression for a total of 10831 genes downloadable from [23] is used in this experiment. For each gene, there are ten expression values, five from a group of insulin-sensitive (IS) people and five from a group of insulin-resistant (IR) people. Only the genes that have no null expression values are included in this analysis. We also require that, for a gene to be included, at least five out of its ten expression values are greater than 100. This eliminates the genes whose expression values are noisy and not reliable.

The results of CM-test are compared with the results of t-test and rank sum test. As we can seen in Table II, although the orders of ranking are different for different methods, all three methods identify these genes as significantly differentially expressed between the IS and IR groups. Furthermore, 10 worst ranked genes in CM-test shown in Table II are also consistent with the result of the other two methods. However, gene U49835 is identified by CM-test as the 22nd ranked significant gene with p-value 0.018, neither t-test (with p-value 0.0768) nor rank sum test (with a p-value 0.1522) identifies this gene as significant. Human chondrocyte protein 39 (YKL-39, gene U49835) or chitinase 2-like protein 2 is up regulated in osteoarthritic chondrocytes. Recently, it has also been shown that patients with osteoarthritic or rheumatoid joint disease have autoimmunity against YKL-39 [13], amine, in combination with chondroitin sulfate, is a popular natural supplement used widely to treat osteoarthritis. However, the use of glucosamine has been linked to development of insulin resistance [21]. No direct evidence of YKL-39 has been reported in insulin resistance cases.

To study the relevance of genes in Insulin metabolism and diabetes, the 10 best ranked differentially regulated genes shown in Table II are further searched in published literature. Human phosphatidylinositol (4,5)bisphosphate 5-phosphatase homolog (gene U45973) was found to be differentially expressed in insulin resistance cases. Over-expression of inositol polyphosphate 5-phosphatase-2 SHIP2 has been shown to inhibit insulin-stimulated phosphoinositide 3-kinase (PI3K) dependent signaling events. Analysis of diabetic human subjects has revealed an association between SHIP2 gene polymorphism and type 2 diabetes mellitus. Also knockout mouse studies have shown that SHIP2 is a significant therapeutic target for the treatment of type-2 diabetes as well as obesity [7]. Csermely et al. reported that insulin mediates phosphorylation/dephosphorylation of nucleolar protein nucleolin (gene $M60858$) by stimulating casein kinase II, and this may play a role in the simultaneous enhancement in RNA efflux from isolated, intact cell nuclei [3]. Schottelndreier et al. [20]

have described a regulatory role of Mx1 and integrin alpha 6 (gene $X53586$) respectively in Ca2+ signaling, which is known to have a significant role in insulin resistance [16]. c-myc is an oncogene that codes for transcription factor Myc that along with other binding partners such as MAX plays an important role widely studied in various physiological processes including tumor growth in different cancers. Myc modulates the expression of hepatic genes and counteracts the obesity and insulin resistance induced by a high-fat diet in transgenic mice overexpressing c-myc in liver [18]. MXI1 (gene L07648) competes for MAX thus negatively regulates MYC function and may play a role in insulin resistance. In the presence of glucose or glucose and insulin, leucine is utilized more efficiently as a precursor for lipid biosynthesis by adipose tissue. It has been shown that during the differentiation of 3T3-L1 fibroblasts to adipocytes, the rate of lipid biosynthesis from leucine increases at least 30-fold and the specific activity of 3-hydroxy-3-methylglutaryl-CoA lyase (gene $L07033$), the mitochondrial enzyme catalyzing the terminal reaction in the leucine degradation pathway, increases 4-fold during differentiation [8]. HCGV gene product (gene $X81003$) is known to inhibit the activity of protein phosphatase-1, which is involved in diverse signaling pathways including insulin signaling [25]. Kobayashi et al. reported that the noradrenaline enhances aortic contractile response in insulin-treated diabetic rats. They also suggested that the insulin deficiency and chronic hyperinsulinemia in diabetes upregulate the IGF-1 receptor and downregulate IGFBP-4 and IGFBP-5 (gene $L27559$) in the aorta. This may be a major cause of the increased vascular contractility resulting in hypertension by hyperinsulinemia in established diabetes [15]. 150-kDa oxygen-regulated protein (ORP150) (gene $U65785$) is a molecular chaperone located in the ER. Studies have shown that ORP150 enhances glucose uptake and simultaneously suppresses oxidized protein in mice. Ozawa et al. reported that, ORP150 enhances the insulin sensitivity of myoblast cells treated with hydrogen peroxide, suggesting that ORP150 is significant for insulin sensitivity, and is a possible candidate for the treatment of diabetes [17].

In summary, out of the top 10 genes identified by CM-test, we could find 8 of them in published literature about their association with insulin metabolism and diabetes. The remaining two genes, $M95610$ and $M88461$, could serve as candidate genes for future research in this area.

V. Conclusions and Future Work

We proposed an innovative approach based on the fuzzy set theory, CM-test, that quantifies the divergence of two sets directly. We have validated CM-test on synthetic datasets and show that it is effective and robust. We also applied CM-test to a real diabetes dataset and identified 10 significant genes. While eight of them have been confirmed to be associated with diabetes in the literature, the remaining two genes, M95610 and M88461, are suggested as two potential diabetic genes for further biological investigation. Further investigation is needed to identify the properties of d-distribution and the precise formula to calculate its p-value.

References

[1] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, NY, 1981.

[2] D. Brown. Classification and boundary vagueness in mapping presettlement forest types. *International Journal of Geographical Information Science*, 12:105–129.

[3] P. Csermely, T. Schnaider, B. Cheatham, M. Olson, and C. Kahn. Insulin induces the phosphorylation of nucleolin. a possible mechanism of insulin-induced rna efflux from nuclei. *J Biol Chem*, 268(13):9747–52, 1993.

[4] A. Davison and D. Hinkley. *Bootstrap methods and their application*. Cambridge University Press, Cambridge, 1997.

[5] J. C. Dunn. Well-separated clusters and optimal fuzzy partitions. *J. Cybern.*, 4-1:95–104.

[6] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3:32–57, 1973.

[7] J. Dyson, A. Kong, F. Wiradjaja, M. Astle, R. Gurung, and C. Mitchell. The SH2 domain containing inositol polyphosphate 5-phosphatase-2: SHIP2. *Int J Biochem Cell Biol*, 37(11):2260–5, 2005.

[8] F. Frerman, J. Sabran, J. Taylor, and S. Grossberg. Leucine catabolism during the differentiation of 3t3-l1 cells. expression of a mitochondrial enzyme system. *J Biol Chem*, 258(11):7087–93, 1983.

[9] I. Gath and A. J. C. Dunn. Unsupervised optimal fuzzy clustering. *IEEE T-PAMI*, 11(7):773–781, 1989.

[10] E. Gustafson and W. Kessel. Fuzzy clustring with a fuzzy covariance matrix. In *Proc. of IEEE CDC*, pages 761–766, Piscataway, NJ, USA, 1979.

[11] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145, 2001.

[12] F. Hoppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis*. Wiley, 1999.

[13] J. T. J, K. Masuko-Hongo, T. Kato, M. Sakata, H. Nakamura, and T. Sekine. Autoimmunity against ykl-39, a human cartilage derived protein, in patients with osteoarthritis. *J Rheumatol*, page 145966, 2002.

[14] G. J. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice-Hall, Upper Saddle River, CA, 1995.

[15] T. Kobayashi, A. Kaneda, and K. Kamata. Possible involvement of igf-1 receptor and igf-binding protein in insulin-induced enhancement of noradrenaline response in diabetic rat aorta. *Br J Pharmacol*, 140(2):285–94, 2003.

[16] R. Kulkarni, M. Roper, G. Dahlgren, D. Shih, L. Kauri, J. Peters, M. Stoffel, and R. Kennedy. Islet secretory defect in insulin receptor substrate 1 null mice is linked with reduced calcium signaling and expression of sarco(endo)plasmic reticulum ca2+-atpase (serca)-2b and -3. *Diabetes*, 53(6):1517–25, 2004.

[17] K. Ozawa, M. Miyazaki, M. Matsuhisa, K. Takano, Y. Nakatani, M. Hatazaki, T. Tamatani, K. Yamagata, J. Miyagawa, Y. Kitao, O. Hori, Y. Yamasaki, and S. Ogawa. The endoplasmic reticulum chaperone improves insulin resistance in type 2 diabetes. *Diabetes*, 54(3):657–63, 2005.

[18] E. Riu, T. Ferre, A. Hidalgo, A. Mas, S. Franckhauser, P. Otaegui, and F. Bosch. Overexpression of c-myc in the liver prevents obesity and insulin resistence. *FASEB J.*, 17(12):1715–7, Sep 2003.

[19] B. Rosner. *Fundamentals of Biostatistics*. Duxbury Press, Pacific Grove, CA, fifth edition, 2000.

[20] H. Schottelndreier, B. Potter, G. Mayr, and A. Guse. Mechanisms involved in alpha6beta1-integrin-mediated ca(2+) signalling. *Cell Signal*, 13(12):895–9, 2001.

[21] M. Wallis, M. Smith, C. Kolka, L. Zhang, S. Richards, S. Rattigan, and M. Clark. Acute glucosamine-induced insulin resistance in muscle in vivo is associated with impaired capillary recruitment. *Diabetologia*, pages 2131–9, 2005.

[22] X. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–846, 1991.

[23] X. Yang, R. Pratley, S. Tokraks, C. Bogardus, and P. Permana. Microarray profiling of skeletal muscle tissues from equally obese, non-diabetic insulin-sensitive and insulin-resistant pima indians. *Diabetologia*, 45:1584?593, 2002.

[24] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

[25] J. Zhang, L. Zhang, S. Zhao, and E. Lee. Identification and characterization of the human hcg v gene product as a novel inhibitor of protein phosphatase-1. *Biochemistry*, 37(47):16728–34, 1998.