

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

This work was supported by the Agricultural Experiment Station at the University of the District of Columbia

Abstract

Problem: 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.

Methods: we propose an innovative approach based on fuzzy set theory, the Cluster Misclassification test (CM-test), to quantify the divergence directly and robustly.

Results: we applied CM-test to diabetes differential gene expression data analysis and identified ten 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.

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.

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 (see Table II), eight of which have been known to be associated with diabetes, with the remaining two genes suggested for further biological investigation.

Methodology

In this section, based on the fuzzy set theory [8], we present our innovative approach, the Cluster Misclassification test (CM-test), to quantify the divergence of two sets of values directly and robustly.

Let $S1$ and $S2$ 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 $S1$ and of $S2$ respectively.
- 2) Characterize $S1$ and $S2$ as two fuzzy sets $FS1$ and $FS2$ whose fuzzy membership functions are defined with the sample means and standard deviations. The fuzzy membership function maps each value x to a fuzzy membership value that reflects the degree of x belonging to the fuzzy set.
- 3) Using the two fuzzy membership functions to 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.

The sample mean of $S1$ is calculated as

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

The sample standard deviation of $S1$ is calculated as

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

We then characterize set $S1$ by a fuzzy set whose fuzzy membership function is defined

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

as 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 $S1$, we measure the degree that it belongs to $FS1$ 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 $FS2$. We say an element in $S1$ is misclassified if it belongs more to $FS2$ 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 $S1$ and $S2$. First, we define the notion of element misclassification degree.

Definition (Element misclassification degree): Given two sets $S1$ and $S2$ and their corresponding fuzzy set $FS1$ and $FS2$, the element misclassification degree of an element e in $S1$ with respect to $FS2$ is defined as

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

We then define the convergence degree of two $S1$ and $S2$ as a linear interpolation of two terms: the number of misclassified elements and the mutual misclassification degrees. Finally, the CM-test d-value of the two sets are defined as 1 minus the convergence degree.

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

where

$$T_1 = \frac{\#M(S_1, S_2)}{|S_1| + |S_2|}$$

$$T_2 = \frac{\sum_{e \in S_1} m(e, S_2) + \sum_{e \in S_2} m(e, S_1)}{|S_1| + |S_2|}$$

$$d(S_1, S_2) = 1 - c(S_1, S_2)$$

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. We also conducted CM-test on a real microarray dataset of diabetes gene expressions to identify genes that are related to diabetes and insulin metabolism.

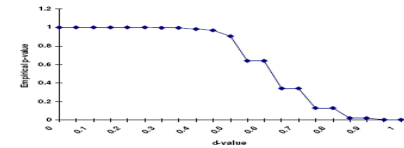


Fig. 1. The relationship between CM d-value and empirical p-value

Suppose two sets $S1$ and $S2$ 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?

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 $S1$ and $S2$ 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 $D \geq 0.8136$, $p\text{-value} \leq 0.05$. Therefore, given two sets $S1$ and $S2$ 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 be considered as significantly divergent.

To apply CM-test to differential gene expression data analysis, a dataset of Microarray gene expression for a total of 10831 genes 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 see 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 FM-test as the 21st ranked significant gene with p-value 0.0258, neither t-test (with p-value 0.0768) nor rank sum test (with a p-value 0.1522) identifies this gene as significant.

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.

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
U45973	750 550 1640 685 636	310 350 135 97 178	1	0.005	0.00	0.008
M60888	301 370 268 323 480	374 506 416 468 440	1	0.005	0.020	0.008
M95610	234 232 215 201 231	82 150 132 282 146	0.910	0.008	0.003	0.015
M88461	218 216 191 181 252	216 200 221 221 219	0.998	0.01	0.01	0.015
S 558	423 095 481 141 242	200 206 229 234	0.991	0.018	0.077	0.152

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
M60888	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 NXI1 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
X81063	Human sapiens HCC 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
X55054	ribosomal protein L23	0.361	0.996	0.272	0.561
X57766	matrix metalloproteinase 11 (stromelysin 3)	0.361	0.996	0.693	0.582
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
L34875	olfactory receptor, family 2, subfamily H, member 2	0.360	0.998	0.524	0.603
M69066	moesin	0.259	0.998	0.403	0.839
M97935	signal transducer and activator of transcription 1, 91kDa	0.359	0.998	0.373	0.839
L20848	oncocadherin	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 (Mow34 homolog)	0.268	1	0.543	0.421