New Brainlike Sensing Solutions for Identifying Deviant and High Impact Data[1]

Overview

Brainlike, Inc. has developed two new computing tools that transform data-based operations into a new era. Conventional data-based operations require that data be gathered in a sample and then analyzed off-line by data analysts. Deviant and high impact data points can be identified by conventional operations, but only as part of a tedious and highly analytic process. Computer programs can also be written to identify deviant or high-impact measurements in real time, using conventional analysis. Again, however, conventional operations require that data must first be gathered and then analyzed off-line by experts. By sharp contrast, the tools described in this report learn from data continuously and automatically, *during* the data gathering process. In the same process and at the same time, they identify deviant or high-impact measurements at once so that appropriate action can be initiated immediately. Moreover, no tedious, highly analytic off-line operations are necessary.

Using these new tools to identify deviant or high impact data automatically and quickly is particularly useful in military, engineering, and health applications. In these and many other applications, reducing the time and effort needed to identify significant new developments can add great value. These tools are especially valuable in settings where information is being produced so quickly, and its nature is changing so rapidly, that the conventional data analysis process cannot keep up.

The first, simple deviance detection (SDD) tool described in this report identifies deviant input values, automatically, adaptively, and in real time. As the SDD program identifies these deviant values, it immediately produces alerts. These alerts, in turn, can trigger on-the-spot actions ranging from immediate preventive activity in surveillance settings to immediate remediation during Internet or telephone surveys.

The second, correlational impact analysis (CIA) tool evaluates the correlational impact of every measured data point. The CIA tool answers the following "five Ws:"

- Who (not necessarily people) in the sample impacts correlations among items the most (either positively or negatively).
- What treatments (group categories, etc.) have the most impact.
- Where in a (not necessarily geographical) sampling region results have the most impact.

[1] Brainlike, Inc. wishes to thank the co-authors of an abstract entitled, "Analytical Innovations To Accelerate Active Living Research Results," by Robert J. Jannarone, Kelli Cain, James F. Sallis, Lawrence D. Frank, Brian E. Saelens, and Harold W. Kohl. Their collaboration in preparing this abstract, which will be published in the Proceedings of the 2004 Active Living Research Conference, has been very useful in demonstrating the need for the tools described in this white paper.

- When (either in time or in some other sequence) during the sampling process results have the most impact.
- Why some individuals, treatments, regions, or periods have more impact than others.

By contrast, the SDD tool answers the following deviance-related questions:

- Who in the sample had the most deviant global, missing, or bias values.
- What measurements in the sample were most deviant.
- Where in a sampling region the most deviance occurred.
- When during the sampling process the most deviance occurred.
- Why some data points were more deviant than others.

Both of these tools are offered as either off-line or real-time computer programs. These programs will be demonstrated and described in the remainder of this report.

Demonstration

		Missing Value Color Code:												
		Extremely Deviant Color Code:					Extremely Deviant Cutoff Value:					3		
		Highly Deviant Color Code:					Highly Deviant Cutoff Value:					2		
	Testee Deviance Values				Testee by Item Deviance Values:									
Testee Labels	Missing Deviance Values	Bias Deviance Values	Global Deviance Values	Item Labels:	A1	A2	B1	B2	B3	C1	C2	C3	C4	C5
1	-0.43	-0.74	1.17		0.34	1.80	1.72	0.20	1.61	0.77	1.18	1.72	1.48	1.37
2	-0.43	1.81	-0.43		0.34	0.44	1.16	1.22	0.51	0.61	1.28	0.55	0.94	0.26
3	-0.43	0.19	-0.22		0.34	0.93	1.16	1.22	0.51	0.61	0.67	0.55	0.34	0.26
4	-0.43	1.28	-0.32		1.07	0.44	0.20	0.82	0.51	0.61	0.05	0.55	0.34	0.26
5	2.24	-1.36	-0.52		0.34	0.93		0.20	1.57	0.08	0.57	0.55	0.27	0.26
6	-0.43	-0.31	-1.39		0.34	0.93	0.20	0.82	0.55	0.08	0.05	0.55	0.27	0.26
7	-0.43	-0.24	0.39		1.07	0.44	0.20	0.82	-4.64	1.46	0.57	0.21	2.08	0.26
8	2.24	-0.15	0.05		0.34	0.93	1.16	0.20	0.55	1.30	1.28		0.94	0.26
9	2.24	-0.17	-0.41		1.75	0.93	0.20	0.82	0.51	0.61	1.28		0.94	0.26
10	2.24	0.53	-0.78		1.07	0.44	0.76	0.20	0.55	0.08	0.05		0.94	0.26
11	-0.43	1.31	2.21		1.07	0.44	1.72	1.22	0.51	0.08	1.80	0.21	2.08	0.26
12	-0.43	-0.22	0.54		1.07	0.44	0.76	0.82	0.55	1.46	1.18	1.72	0.34	0.26
13	-0.43	0.82	-0.62		1.07	0.44	1.16	1.22	1.57	0.61	0.57	0.21	0.34	0.26
14	-0.43	1.98	0.53		0.34	0.93	1.16	1.22	0.51	1.30	1.28	1.31	0.94	0.26
15	-0.43	-1.20	4.97		2.48	3.16	5.02	0.82	1.61	2.15	2.42	2.48	0.87	0.26
16	-0.43	0.95	1.42		1.75	0.93	1.16	2.25	1.57	1.30	0.67	1.31	0.94	0.26
25	-0.43	-0.66	0.20		1.07	0.44	1.16	0.20	0.51	2.15	1.18	1.72	0.27	0.26
		Item Missing Deviance Values:			-0.32	-0.32	0.97	-0.32	-0.32	-0.32	-0.32	3.55	-0.32	-0.32

Figure 1. Simple Deviance Detection Program Output

Figure 1 illustrates results from the SDD program.[2] Deviance values shown in the figure are standardized, so that less deviant observations have values near zero and more deviant observations have higher positive or negative values. Highly deviant values are color coded as yellow and extremely deviant values are color coded as red. Deviance cutoff values are configurable. Cutoff values shown are set to plus or minus 2.0 for the highly deviant category and to plus or minus 3.0 for the extremely deviant category. Missing values are color coded as magenta.

Each row in Figure 1 represents a distinct "testee." Testees may be persons, time points, or any other units for which data are available. Each column in the figure represents a distinct, measured "item." Each item has either a deviance value or a missing value for each testee. Testee by Item Deviance values indicate how unexpected each item value is for each testee. Missing Deviance Values indicate the deviance level of missing value counts for each testee as well as for each item. Bias Deviance Values indicate how much testees tended to have either high or low average deviance values over all items. Global Deviance Values indicate how much testees tended to either high or low average deviance magnitudes over all items.

One red value in Figure 1 shows that testee number 15 had extremely deviance values overall, and that the main sources were items A2 and B1. This testee's deviant item scores could have been due to unusually high paired values or low values paired values for those items. If this result had been available to the surveyor during an interview or to the testee during an Internet survey, the surveyor or testee could have re-examined the result and corrected errors on the spot. Another deviance value in Figure 1 shows that item C3 had an extremely high number of missing values. Had this result been found at the beginning of a survey, the item could have been re-examined for clarity at once and corrected accordingly.

Results like those in Figure 1 can be generated and displayed on a row by row basis by an auto-adaptive version of the SDD program, so that immediate actions can be taken in real-time surveying or surveillance settings. In such settings, the program receives one new row of data at each time point. After reading the row data, the SDD program displays the corresponding deviance values for that row, along with updated Item Missing Deviance Values. The results in Figure 1 can also be generated and displayed all at once, after sampling all rows of data have been observed and supplied to an off-line version of the SDD program.

Figure 2 illustrates results from the correlational impact analysis program. The values in Figure 2 are standardized and color coded as in the SDD case, but in the CIA case they measure contributions to correlation impact rather than to deviance. Each Testee by Item impact value indicates one testee's contribution to the multiple correlation coefficient for a particular item. For example, the large negative value for testee 5 and item B3 in Figure 2 indicates that the multiple correlation coefficient for predicting B3 from all other items would be much higher if testee 5 were excluded. Likewise, the large positive

[2] Figure 1 shows more significant deviations than would be expected from actual data — input and output results using actual data can be obtained by contacting Brainlike, Inc..

value for testee 15 and item A2 indicates that the multiple correlation coefficient for predicting A2 from all other items would be much lower if testee 15 were excluded.

	Extremely Deviant Color Code:						Extremely Deviant Cutoff Value:					3
	Highly Deviant Color Code:						Highly Deviant Cutoff Value:					2
			Testee by Item Deviance Values:									
Testee Label	Testee Deviance Values	Item Labels:	A1	A2	B1	B2	B3	C1	C2	C3	C4	C5
1	1.17		-2.14	1.80	1.72	-0.21	-1.61	0.77	-0.51	2.48	1.48	-0.02
2	-0.43		1.22	0.44	1.16	0.74	0.05	2.07	1.57	1.07	-0.94	0.24
3	-0.22		0.82	0.93	-1.16	-1.20	0.57	0.34	0.55	-0.34	2.13	0.07
4	-0.32		2.25	0.44	0.20	-0.23	1.28	-0.34	2.65	-1.75	0.34	-0.31
5	-2.94		0.55	0.93	0.92	-0.08	-3.12	-2.57	0.57	1.07	0.27	0.03
6	-1.39		0.34	0.93	0.20	-1.72	0.55	0.08	-0.05	1.07	0.27	0.00
7	0.39		-1.07	0.44	-0.20	0.21	0.51	2.22	-0.57	-0.34	-1.57	-0.24
8	0.05		1.46	0.93	1.16	1.31	0.55	1.22	1.28	0.34	-0.94	0.09
9	-0.41		0.61	-0.32	-0.32	-0.32	0.20	1.32	2.16	-0.34	-0.94	0.07
10	-0.78		1.30	0.44	0.76	0.20	0.55	1.89	0.05	-1.07	-0.94	-0.11
11	1.93		2.15	0.44	1.72	-2.63	0.51	0.82	2.17	0.34	-2.08	0.32
12	0.54		1.07	0.44	0.76	0.82	0.55	0.82	1.18	-1.18	0.34	-0.04
13	-0.62		-1.07	0.44	-1.16	1.22	1.57	-0.61	0.57	1.28	-0.34	0.13
14	0.53		0.34	0.93	1.16	-1.22	0.51	1.30	1.28	-0.67	0.94	-0.02
15	4.23		2.48	7.82	3.94	0.82	1.61	2.15	-0.42	-0.20	-1.17	0.39
16	1.42		1.75	0.93	-1.16	-0.73	1.57	-1.30	0.67	-1.31	0.94	0.02
25	0.20		0.93	1.16	-1.31	0.55	1.22	-1.28	0.34	0.94	-0.09	0.06
	Item Deviance Values:		-0.32	5.17	3.28	-1.60	-2.81	-2.09	1.74	0.51	0.82	-3.82

Figure 2. Correlational Deviance Detection Program Output

Testee Deviance Values in Figure 2 indicate how much individuals tended to have either high or low average deviance values over all items. Item Deviance Values indicate whether each item's average contribution to the other items' multiple was low or high. For example, testee 15 contributed extremely positively to inter-item correlation values. The pattern shown for testee 15 and the item Deviance values for A2 and B1 would occur if a paired outlier value appeared in the upper right or lower left quadrant of a standard scatter plot between items A2 and B1. By contrast, the deviance for testee 5 contributed extremely negatively to inter-item correlation coefficient values. The pattern showed for testee 5 and the Item Deviance values for B2 and B3 would occur if a paired outlier value appeared in the lower right or upper left quadrant of a standard scatter plot between items B2 and B3. The Item Impact value for C5 and its pattern of Testee by Item Impact Values is consistent with an item that has conspicuously low predictive value.

In the process of developing these tools, useful solutions have been discovered to a variety of important practical problems. These include performing all the necessary computing operations very quickly, automatically and efficiently detecting and replacing missing or deviant data, and getting useful information out small-sample data. To give one example, the CIA tool produces useful comparative impact results for testees and items after learning from only a small number of testees, even when the number of items greatly exceeds the number of testees.

Statistical Details

For the SDD program, each column of Item by Testee Deviance values in Figure 1 is standardized with respect to its corresponding column mean and standard deviation. Row and column means and standard deviations that produce these deviance values are based on corresponding measured values that are non-missing. Testee Missing Deviance Values in the figure are standardized counts of missing values for each testee, relative to all others. Item Missing Deviance Values are computed similarly. Each Bias Deviance value is obtained by averaging all non-missing Testee by Item Deviance values in its corresponding row and then standardizing over all such average values. Each Global Deviance value is obtained by averaging the magnitude of all non-missing Testee by Item Deviance values in its corresponding row and then standardizing over all such average values.

For the CIA program, each Testee by Item Deviance value is the standardized difference between (a) the multiple correlation coefficient for predicting its corresponding column item based on all other column items with that testee included, and (b) that same multiple correlation coefficient with that testee excluded. Each Testee Deviance value is obtained by averaging all non-missing Testee by Item Deviance values in its corresponding row and then standardizing over all such average values. Each Item Deviance value is the standardized difference between (a) the average among multiple correlation coefficients for predicting all other items with that item included, and (b) that same average with that item excluded.

In the auto-adaptive version of the SDD and CIA programs, means and standard deviations are retrieved by the program each time a row of data arrives, and they used to computed deviance values. They are then updated recursively for future use. The CIA program, also uses and updates correlation coefficients similarly. Missing values are excluded from updating means, standard deviations, and correlation coefficients. Optionally, extremely deviant values are excluded as well. In addition, the CIA program uses a sophisticated updating algorithm to update its parameters efficiently. Otherwise, computing would take far too much time to be practical in settings involving many variables. The CIA algorithm also uses advanced techniques to handle items that are linearly redundant, as in cases where the number of items is large relative to the number of testees.

Summary

This report has introduced two new computing tools that Brainlike, Inc. has recently developed. One tool identifies deviant input values and the other identifies high impact input values. Both tools operate automatically, adaptively, and in real time. Both tools offer high value for accelerating and automating the data analysis process, to the point that actions can be taken based on deviant and high impact items on the spot, without the need for manual statistical analysis.