International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:7, 2010

Dataset Analysis Using Membership-Deviation Graph

Itgel Bayarsaikhan, Jimin Lee, and Sejong Oh

***Abstract*—**Classification is one of the primary themes in computational biology. The accuracy of classification strongly depends on quality of a dataset, and we need some method to evaluate this quality. In this paper, we propose a new graphical analysis method using 'Membership-Deviation Graph (MDG)' for analyzing quality of a dataset. MDG represents degree of membership and deviations for instances of a class in the dataset. The result of MDG analysis is used for understanding specific feature and for selecting best feature for classification.

***Keywords*—** feature, classification, machine learning algorithm.

I. Introduction

CLASSIFICATION is one of the primary themes in computational biology and bioinformatics. In service of this task, we select a specific *dataset*, and then train *classifiers* using known classified data. The accuracy of training and classification depends on the quality of the dataset. If a dataset has separable classes, the classification accuracy of the feature may be good. But it is difficult to measure the separability of a dataset.

Feature selection [1-6] is the problem of selecting the best feature that is, ideally, necessary and sufficient to describe the target concept [7]. The objective of feature selection is to obtain a dataset characterized by (1) low dimensionality, (2) retention of sufficient information, (3) enhancement of separability in feature space for examples in different categories via the removal of effects resulting from noisy features, and (4) the comparability of features among examples in the same category [8]. Most of feature selection algorithms have an evaluation function that produces scores for candidate datasets. Let us suppose that F_1 and F_2 are datasets and $E(x)$ is an evaluation function, and if $E(F_1) > E(F_2)$, we can expect that dataset F_1 will yield better training/testing accuracy than F_2.

We can use the evaluation functions to understand and analyze specific dataset. We introduce some evaluation functions in section 2. The evaluation functions produce numerical scores, and we guess the characteristics of a dataset depending on the numerical values. If we can visualize the vales, we can easily understand characteristics of a dataset.

In this paper, we propose a graphical analysis method, called 'Membership-Deviation Graph (MDG)'. MDG is a scatter diagram for a class of a dataset, and based on two evaluation values, degree of membership and deviations for instances in the class. From the MDG, we can observe the distributions and accuracies of each class in a dataset. It means that we can decide about the quality of the feature.

The remainder of this paper is structured as follows. Section 2 summarizes several dataset evaluation functions. Section 3 describes the proposed Membership-Deviation Graph. Section 4 describes the results of experiments concerning MDG. We apply MDG on Yeast dataset, and the conclusions of this paper are provided in Section 5.

II. Dataset Evaluation Functions

The goal of dataset evaluation is to calculate the distances among classes in the dataset. Let D be a distance function and $D(C_1,C_2) > D(C_3,C_4)$, thus making classes C_1 and C_2 more separable than C_3 and C_4.

The Euclidean distance is a base of other distance functions. The Euclidean distance between the two points $p(p_1,p_2,p_3,..,p_n)$ and $q(q_1,q_2,q_3,..,q_n)$ is defined as follows :

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + .. + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (1)$$

The Hausdorff distance [9] measures how far two subsets of a metric space are from one another. The Hausdorff distance between set A and set B is defined as follows:

$$H(A, B) = max\{h(A, B), h(B, A)\} \quad (2)$$

In the equation (2), The directed function $h(A, B)$ refers to the distance between set A and set B, and $H(A, B)$ is also the distance between sets A and B.

J. Liang *et al.* [10] previously suggested a new feature selection algorithm based on a distance discriminant (FSDD). They used the distance among center points of categories, the standard deviation of each attribute in a feature, and the intra-set distance of categories.

RELIEF [11-13] is regarded as one of the more successful feature selection algorithms. The basic idea of RELIEF is to iteratively estimate feature weights according to their ability to discriminate between neighboring instances. In each iteration, an instance x is selected randomly, and then the

Itgel Bayarsaikhan., with the WCU Research Center of NanoBioMedical Science, Dankook University, Korea (e-mail: superitgel@yahoo.com).
Jimin Lee is with the Department of Computer Science, Dankook University, Korea (e-mail: may-3rd@cyworld.com).
Sejong Oh is with the WCU Research Center of NanoBioMedical Science, Dankook University, Korea(corresponding author to provide phone: 82-41-550-3484; fax: 82-41-550-1149; e-mail: sejongoh@ dankook.ac.kr).

nearest instance of it is found from the same class (NH) as well as different classes (NM). Finally, the weight value is updated by the equation:

$$w_i = w_i + |x^{(i)} - NM^{(i)}(x)| - | x^{(i)} - NH^{(i)}(x) | \quad (3)$$

As we mentioned before, above evaluation functions produce numerical values, and it is difficult to imagine the characteristics of specific feature. In the next section, we describe a new visual analysis tool, Membership-Deviation Graph.

III. Membership-Deviation Graph

A. Summary

MDG is a two-dimensional scatter diagram as shown in Fig.1. One diagram contains information for a class in a dataset, so we may generate 7 diagrams if a dataset has 7 classes. Each point on the diagram corresponds to instance of represented class. In the diagram, the axis of *X* represents the deviation of distance from center point (instance) of a class, and the axis of *Y* represents the degree of membership of each point into the class.

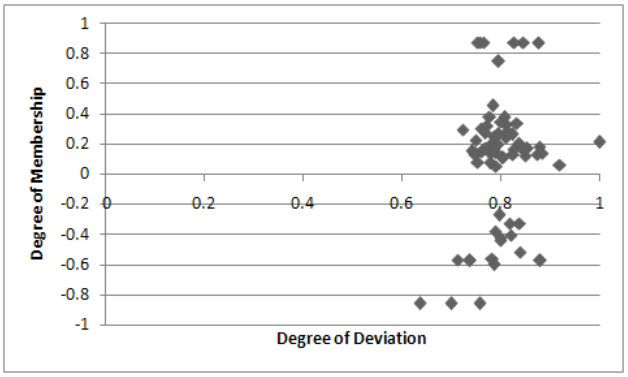

Fig. 1 An example of MDG

Fig. 2 shows the meaning of deviation. The scope of deviation of *Case A* is narrower than *Case B*. It means that the points in *Case B* are more widely scattered than one in *Case A*. Fig. 3 shows the meaning of membership. The membership of points in *Case C* is stronger than one in *Case D*. If instances in a class have strong membership and narrow scope of deviation, the class may be well-separable from other classes. We describe more detail meaning of deviation and membership in the next parts.

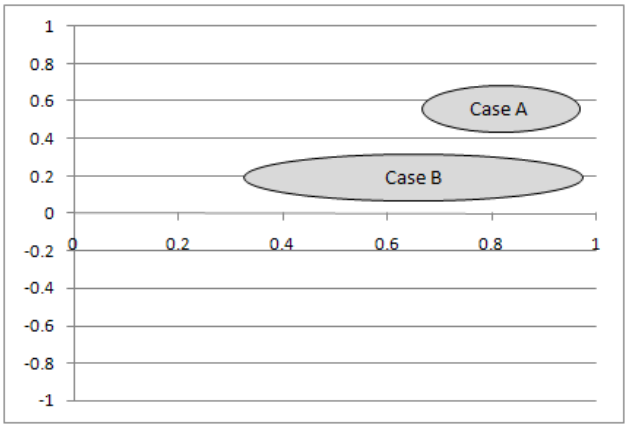

Fig. 2 Two cases of deviation

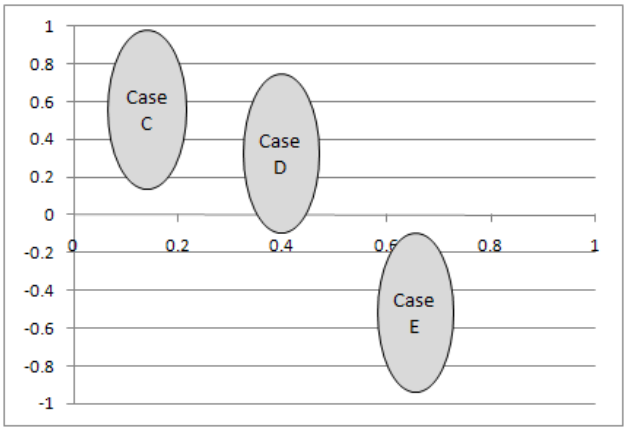

Fig. 3 Three cases of membership

B. Degree of Membership

A class may have many instances, but the memberships of each instance are different. In Fig. 4, for example, P_1 and P_2 are belongs to class C_1, but C_2 is located in overlapping area with C_2. Therefore, no classifier can easily predict P_2 is belongs to class C_1. We calculate degree of membership for an instance by counting the number of nearest instances that belongs to same class. If the degree of membership is 1, it means that the instance is located fully far from other classes, and can be easily classified. If the degree of membership is -1, it means that the instance is located in deeply overlapped area with other classes, and very difficult to correctly classify.

The membership function $M(P_i)$ is defined as follows:

$$M(P_i) = \frac{Cnt\,(P_i)}{\mathrm{K}} \times \frac{AvgD\,(SP_i)}{AvgD\,(\mathrm{SP_i}) + AvgD\,(NP_i)} \quad (4)$$

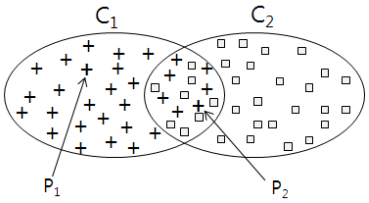

Fig. 4 Two instances in a same class

In the equation (4), K is the total number of nearest instances of instance P_i, and function $Cnt(P_i)$ returns the number of nearest instances that belongs to same class of instance P_i. Therefore, $\frac{Cnt(P_i)}{K}$ means the ratio of homogeneous instances of P_i over whole K-nearest neighbor instances. In the second part of equation (4), SP_i is homogeneous instances of P_i and NP_i is non-homogeneous instances of P_i over whole K-nearest neighbor instances. $AvgD(SP_i)$ is average distance of SP_i. Important thing of $M(P_i)$ is that it has positive value if is correctly classified by KNN classifier, or it has negative value. For simplicity we normalize $M(P_i)$ into [-1,1].

Degree of Deviation

The degree of deviation for an instance has intuitive meaning and we just describe the deviation function $DIV(P_i)$ as follows:

$$DIV(P_i) = ED(P_i, C) \quad (5)$$

In the equation (5), $ED(P_i, C)$ is a function that returns Euclidean distance between P_i and center point of P_i's class. Before drawing MDG, we normalize the output of $DIV(P_i)$ into value range [0,1].

IV. Experiments

A. Dataset

To show usefulness of MDG, we choose Yeast dataset [14], and analyze it using MDG. Original Yeast dataset has 10 classes, but we remove a class that has a little bit instances. TABLE I summarize Yeast data set. We test K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), and Support Vector Machine (SVM) classifiers, and TABLE II summarize the classification accuracies.

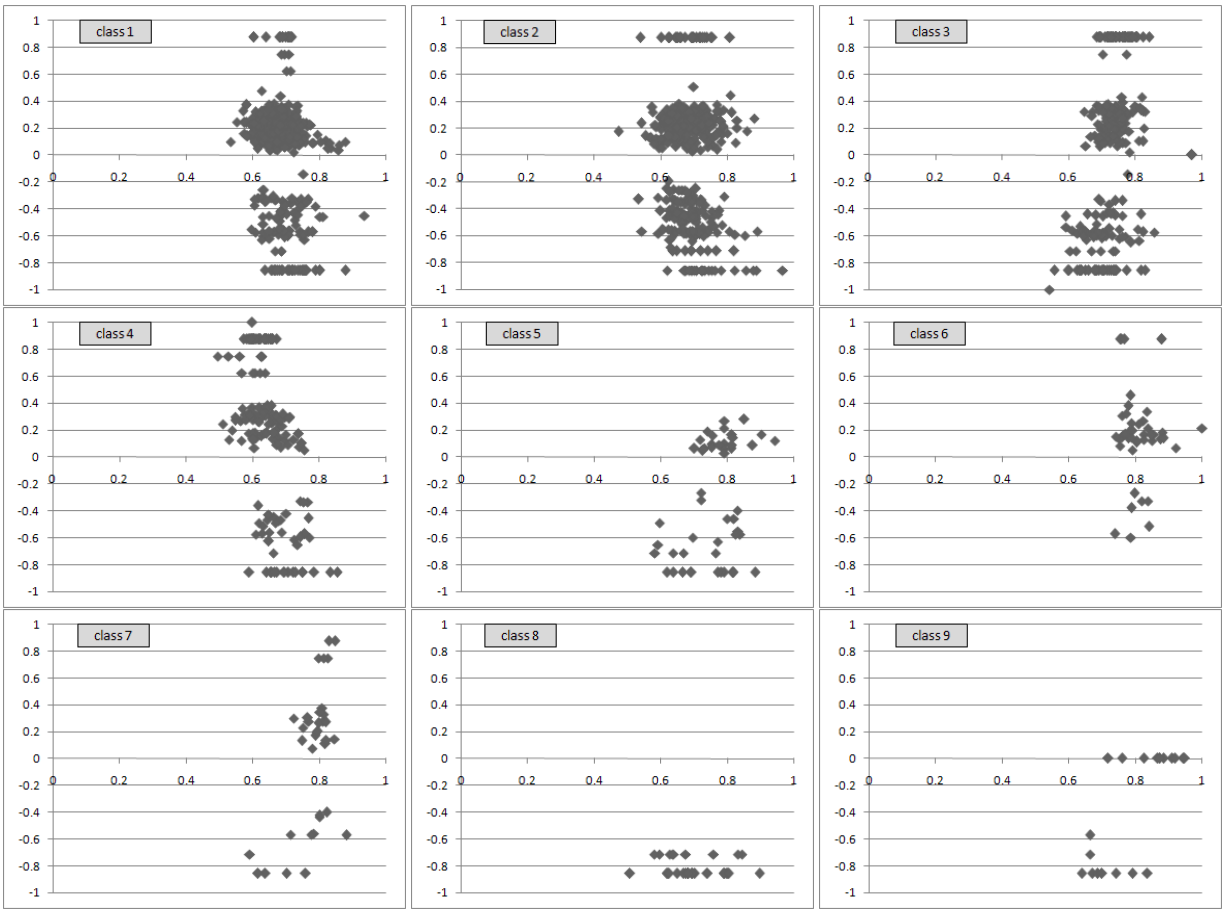

Fig. 5 Result of MDG analysis for Yeast dataset

B. MDG Analysis Results

Fig. 5 summarizes MDG analysis results. In the case of *class 8*, membership degrees of all instances have large negative values; it means that all instances of *class 8* are located in overlapped area with other classes. Therefore, the classification accuracies produced by 3 classifiers for *class 8* are all zero. In the case of *class 6*, the ratio of instances in overlapped area is relatively small and also the scope of deviation is relatively narrow; 3 classifiers produce high classification accuracy. Classification accuracy of *class 1* is greater than *class 2* because *class 2* has more points that have negative membership value.

TABLE I
GENERAL INFORMATION OF YEAST DATASET

No of instances	No of Classes	No of attributes
1484	9	9

TABLE II
CLASSIFICATION ACCURACIES FOR YEAST DATASET

Class	KNN	ANN	SVM
Class 1	0.69	0.16	0.86
Class 2	0.47	0.38	0.3
Class 3	0.53	0.43	0.5
Class 4	0.65	0.66	0.68
Class 5	0.23	0.27	0.0
Class 6	0.82	0.68	0.59
Class 7	0.61	0.39	0.0
Class 8	0.0	0.0	0.0
Class 9	0.6	0.1	0.6
Total	0.57	0.31	0.53

From the MDG, we can see distributions of instances in a class, and we can compare each class in a feature. Now MDG is just observation tool, but if we process the result of MDG, we may be able to get more useful information for feature selection.

V. CONCLUSION

MDG is a scatter diagram to show distributions of instances in a class. It contains information about degree of membership and deviation for an instance in a class. We can observe each class and compare classes of a dataset using MDG. We also can guess the approximate quality of the dataset. Comparison of the MDG results from different datasets is one of further works.

ACKNOWLEDGMENT

This work was supported by grant No. R31-2008-000-10069-0 from the World Class University (WCU) project of the Ministry of Education, Science & Technology (MEST) and the Korea Science and Engineering Foundation (KOSEF) through Dankook University.

REFERENCES

[1] H. Liu, J. Li, L. Wong, "A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns", Gene Informatics 13, 2002, pp51-60.
[2] S. Doraisamy, S. Golzari, N.M. Norowi, M.N.B Sulaiman, N.I. Udzir, "A Study on Feature selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music", Proc. of International Conference on Music Information Retrieval, 2008, pp331-336.
[3] I. Guyon, A. Elisseeff, "An introduction to variable and feature selection", J. Mach. Learn. Res. 3, 2003, pp.1157–1182.
[4] R. Gilad-Bachrac, A. Navot, N. Tishby, "Margin based feature selection—theory and algorithms", Proceedings of the 21st International Conference on Machine Learning, 2004.
[5] K.H. Quah, C. Quek, "MCES: a novel Monte Carlo evaluative selection approach for objective feature selections", IEEE Trans. Neural Networks 18 (2), 2007.
[6] J. Dy, C.E. Brodley, "Feature selection for unsupervised learning", J. Mach. Learn. Res. 5, 2005, pp845–889 2005.
[7] K. Kira, L.A. Rendell, "A Practical Approach to Feature Selection", Proceedings of the Ninth International Conference on Machine Learning, 1992, pp249-256.
[8] W.S. Meisel, Computer-Oriented Approaches to Pattern Recognition, Academic Press, New York, 1972.
[9] S. Piramuthu, "The Housdorff Distance Measure for Feature Selection in Learning Applications", Proceedings of the 32nd Hawaii International Conference on System Sciencespp1-6, 1999.
[10] J. Liang, S. Yang, A. Winstanley, "Invariant Optimal Feature Selection: A Distance Discriminant and Feature Rranking Based Solution", The journal of the pattern recognition, 2008, pp1429-1439.
[11] K. Kira, and L.A. Rendell, "The feature selection problem: Traditional methods and a new algorithm", Proceedings of Ninth National Conference on Artificial Intelligence, 1992, pp129–134.
[12] Y. Sun and D. Wu, "A RELIEF Based Feature Extraction Algorithm", Proceedings of the 2008 SIAM International Conference on Data Mining, 2008, pp188-195.
[13] I. Kononenko, E. Simec, M. Robnik-Sikonja, "Overcoming the myopia of induction learning algorithms with RELIEFF", Applied Intelligence Vol7, 1, 1997, pp.39-55
[14] K. Nakai, Yeast Dataset, http://archive.ics.uci.edu/ml/datasets/Yeast.

Sejong Oh received a Doctor, Master, and Bachelor degree in Computer Science from Sogang University, Korea, in 2001, 1991, and 1989. From 2001 to 2003, he was a postdoctoral fellow in the laboratory for Information Security Technology at George Mason University, USA. Since 2003 he joined the Department of Computer Science at Dankook University, Korea, and is currently associate professor in WCU Research Center of NanoBioMedical Science. His main research interests are bioinformatics, information system, and information system security.