Effectiveness of Value Granulation in Machine Learning for Massively Large and Complex Domain

Zachary I Moore
Computing and Engineering Science
Eastern Washington University
Cheney, WA 99004-2493 USA

Atsushi Inoue *
Computing and Engineering Science
Eastern Washington University
Cheney, WA 99004-2493 USA
E-mail: atsushi.inoue@ewu.edu

Abstract

Considered from data analysis and dynamic optimization view point, computer networks are massively large and complex data domains where analytical computation (e.g. statistics) is, despite its needs, often found infeasible. In an attempt to address the issues raised by such domains, we are currently studying Granular Computing, a newly emerging paradigm, and its application to Machine Learning. This paper reports effectiveness of value granulation (such as discretization and quantization) in Machine Learning from aspects of complexity reduction, learning capability, and intelligent system development.

Keywords: Granular Computing, Machine Learning, Complex Data Domain.

1 Introduction

Considering the nature of massively large size and high dimensionality (aka sparseness) of data extracted from computer networks, Machine Learning and Data Mining approaches are primarily considered. However, many conventional approaches suffer not only from extremeness of those natures but also from its time-sensitive demands and thus result in very limited practicality.

We have been studying applications of Support Vector Machine (SVM) to the data set for DARPA IDS evaluation project at MIT Lincoln Lab[3]. This data set is generated, for the purpose of evaluating intrusion detection systems, in a simulated networking environment with various intrusive cases and consisting of many large and highly dimensional netlogs. Our previous findings are summarized as follows:

Large-Scale SVM Learning We proposed and studied a SVM learning method, so-called ArraySVM[5]. This outperforms other conventional large-scale SVM learning in both training time (two to three times less) and classification accuracy (by 10-20%).

Online SVM Learning We proposed and studied an online learning method that is slightly modified from Sequential Minimal Optimization (SMO) algorithm[4]. Our study showed that this online learning outperforms SVM classifiers trained by conventional SMO by 10% and yields simpler SVM classifiers in terms of the number of support vectors. This enables reinforcement of SVM learning in a timely manner while maintaining quality performance.

To seek for more simplicity and unity in our approaches, we are studying Granular Computing[10, 1] as an underlying scheme of Machine Learning in massively large and complex data domains such as computer networks. First, we applied Granular Com-

* Corresponding author.

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 855–862
Torremolinos (Málaga), June 22–27, 2008

puting to SVM, so-called Granular Support Vector Machine (GSVM), in the same problem setting as the above large-scale SVM learning[7]. Later, we extended this study to a general framework of Machine Learning (binary classification problems)[6]. Our study showed that classification accuracy with Granular Computing scheme outperforms the one without this scheme by 10-25%. We also found that ArraySVM, our most outperforming large-scale SVM learning, is an instance of this framework, and an extension of ArraySVM with information granules based on geometric proximity, so-called Granular Expert SVM (GESVM), was proposed. It was impressive that GESVM outperforms the original ArraySVM by 5-10% and reached to the classification accuracy of 99.5%.

We are now wondering about the cause of outperformace as a result of Granular Computing, in other words how Granular Computing is effective in Machine Learning. In this paper, we first discuss our scheme of constructing and modifying information granules, and then its effectiveness in Machine Learning from aspects of complexity reduction, learning capability, and intelligent system development.

2 Information Granules

In brief, Granular Computing concerns the processing of abstracted information entities, so-called information granules. An information granule is a collection of entities (data points) that usually originate at the numeric level and are arranged together due to their similarity, functional adjacency, indistinguishability, coherency, or the like[10, 1].

2.1 Representation

Representation of information granules may vary depending on the type of granulation. In many cases, we consider one of the following: sets and intervals, rough sets, and fuzzy sets.

Fuzzy sets are capable of representing rough sets, crisp sets, intervals, and values as their spacial cases. When the membership function of a fuzzy set $\mu(x)$ is either 0 or 1, that fuzzy set is crisp. A rough set A is a construct of two crisp sets, namely a lower approximation $\underline{A}$ and an upper approximation $\overline{A}$, in order to approximate a target set T whose members cannot be completely determined by given attributes.

By *Decomposition (aka Representation) Theory*[8], fuzzy set F can be decomposed of (i.e. represented as) a collection of crisp sets F_α (so-called α-cut) such that

$$F = \bigcup_{\alpha \in [0,1]} \alpha \cdot F_\alpha \qquad (1)$$

where $\alpha \cdot F_\alpha$ is a special fuzzy set whose membership function is given by $\mu_{\alpha \cdot F_\alpha}(x) = \alpha$ for all $x \in F_\alpha$.

There is a correspondence between a rough set and a fuzzy set in terms of α-cut. For a target set T and its rough set $A =< \overline{A}, \underline{A} >$, T corresponds to F_{α_0}, $\overline{A}$ corresponds to F_{α_1} and $\underline{A}$ corresponds to F_{α_2} where $\alpha_2 \geq \alpha_0 \geq \alpha_1$ and $\alpha_0, \alpha_1, \alpha_2 \in [0, 1]$.

Consequently, information granules represented by fuzzy sets and rough sets may consist of a collection of corresponding crisp sets.

2.2 Construction

In conventional Granular Computing, there are three types of granulation:

- **Value Granulation** corresponding to discretization and quantization.
- **Variable Granulation** corresponding to clustering, aggregation and transformation.
- **Concept Granulation** corresponding to component analysis.

Obviously, *value granulation is the simplest* as it can be achieved by discretization and quantization. Formally, the following procedure is to be performed:

1. Domain X^n is granulated by values such that each dimension is divided into C_i uniform segments ($1 \leq i \leq n$).

2. Non-empty segments are considered as information granules.

Such information granules are crisp sets and are not overlapped each other. Their set Cardinalities vary. Using set operations, more highly abstracted information granules can be composed and vise versa. Using Decomposition theory, fuzzy sets can be constructed, as well as rough sets.

2.3 Application to Machine Learning

As for applications of Granular Computing to Machine Learning, we are studying two approaches according to Zadeh's granulation model[10, 6]: information granulation and action granulation.

Information granulation first selects a pivotal value per information granule and then applies a particular Machine Learning algorithm to the set of that pivotal value. The selection of that pivotal value is represented as an *admission function*[6]. Among selections of the admission function, we normally use averaged vectors. GSVM is an example of information granulation consisting of a value granulation and an application of SVM[7].

Action Granulation follows the Divide-and-Conquer scheme in such a way that it first applies a particular Machine Learning algorithm to each information granule and then aggregate the results. ArraySVM and GESVM are examples. The former consists of a stochastic variable granulation (uniform distribution of samples over information granules) and an application of SVM[5]. The latter consists of a value granulation and an application of SVM[6].

3 Complexity Reduction

Conceptually, any granulation (value, variable and concept) reduces computational complexity of tasks applied to information granules due to the fact that $|G| \leq |D|$, where G is a set of information granules generated from a sample data set $D = X_1 \times \cdots \times X_n$. The efficiency of this size reduction can be well measured by $|G|/|D|$. Variable and concept granulation is achieved as a result of applying some dynamic analytical process such as clustering, aggregation (mainly stochastic) and component analysis. Consequently, this yields a good reduction, i.e. a small $|G|/|D|$.

On the other hand, value granulation does not necessarily guarantees a small $|G|/|D|$. In the worst case, this may yield $|G| = |D|$ when only one sample constitutes an information granule. However, its construction is significantly advantageous in terms of computational complexity. Most of dynamic analytical processes such as clustering and component analysis take their computational complexity in the order of $|D| \cdot \log |D|$ or $|D|^2$, whereas quantization and discretization take that in the order of $|D|$.

To examine the size reduction of value granulation, we conduct an experiment of simple value granulation applied to a massively large, high dimensional data set. In this experiment, a value granulation is applied to the data set for DARPA IDS evaluation project at MIT Lincoln Lab[3] in order to study efficiency of size reduction. This data set consists of thirty-four continuous features and seven categorical features as follows:

- Basic features of individual TCP connections (9 features: 5 continuous, 4 categorical)
- Content features within a connection suggested by domain knowledge (13 features: 10 continuous, 3 categorical)
- Traffic features computed using a two-second time window (9 features: 9 continuous, 0 categorical)
- Traffic features of the destination host computed using a two-second time window (10 features: 10 continuous, 0 categorical)

There are twenty-two classes as follows:

- normal network activities.

C	$\|S\|$	$\|G\|$	Gsize[min, max, ave]
2	2^9	146	[1, 243, 69]
3	3^9	591	[1, 72, 18]
4	4^9	585	[2, 30, 17]
5	5^9	3253	[1, 9, 3]
10	10^9	7311	[1, 17, 1]

Table 1: Value Granulation ($|D| = 10^4, |S| = C^9$)

C	$\|S\|$	$\|G\|$	Gsize[min, max, ave]
2	2^9	15512	[1, 196, 61]
3	3^9	51259	[5, 97, 18]
4	4^9	67319	[2, 31, 16]
5	5^9	214503	[1, 21, 4]
10	10^9	582122	[1, 231, 17]

Table 2: Value Granulation ($|D| = 10^6, |S| = C^9$)

- network attack or misuse, consisting of four different types: denial of service (6 classes, e.g. neptune), remote to local (8 classes, e.g. guess passwprd), user to root (4 classes, e.g. buffer overflow), and surveillance/probing (4 classes, e.g. portsweep).

The value granulation is specified by a constant C such that each continuous feature X is granulated into C segments in $[\min[X], \max[X]]$. Categorical features are ignored in this experiment. With randomly selected records, we examine the following factors for different data size ($|D|$) and different dimensions.

- The number of segments, i.e. $|S|$.
- The number of information granules, i.e. $|G|$.
- The granule size – minimum, maximum and average.

Tables 1, 2 and 3 show the results with various data size ($|D|$) and dimensionality (i.e. corresponding to the increase of $|S|$). The pair of 'Gsize' in those tables represents the minimum number of records in a granule, the maximum and the average number of records respectively.

C	$\|S\|$	$\|G\|$	Gsize[min, max, ave]
2	2^{19}	754	[1, 37, 11]
3	3^{19}	1035	[1, 33, 10]
4	4^{19}	988	[1, 51, 9]
5	5^{19}	1071	[1, 18, 8]
10	10^{19}	3743	[1, 36, 3]

Table 3: Value Granulation ($|D| = 10^4, |S| = C^{19}$)

4 Learning Capability

We discuss effectiveness of value granulation in Machine Learning in various cases. Effectiveness of value granulation applied to Support Vector Machine (SVM) has been studied in [7] and [5]. Discussions in this section are based on our findings, the cases of SVM. We then propose a simple classifier learning method that utilizes the effectiveness of value granulation.

4.1 Information Granulation vs. No Granulation

We studied SVMLight, a standard and popular SVM optimization tool, that uses a fast optimization with heuristics[2]. This tool is known as one of efficient SVM optimization with thousands of records.

For value granulation, we choose dimension segmentation parameter C to be multiples of ten (e.g. 10, 20, 30, etc.). For each granule, at most two vectors are generated as a result of information granulation: the average vector of positive class and that of negative class. Only one average vector is generated when the granule consists of vectors in one class. SVMLight is then applied to the set of those average vectors, i.e. so-called GSVM.

We conduct experiments in two cases: a thousand of records from the DARPA IDS evaluation data set ($|D| = 1000$) and ten thousands of records from the same data set ($|D| = 10000$). In each case, those records are randomly selected for learning (i.e. SVM optimization), and the other set of the records with the same number is also randomly selected for measuring classification accuracy. We measure the optimization time and the

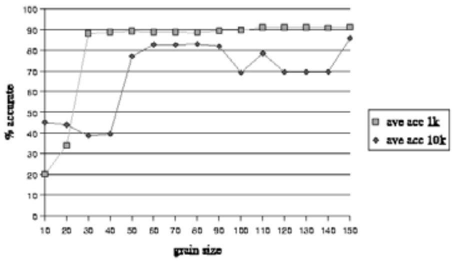

Figure 1: Classification Accuracy ($C = 10^n$)

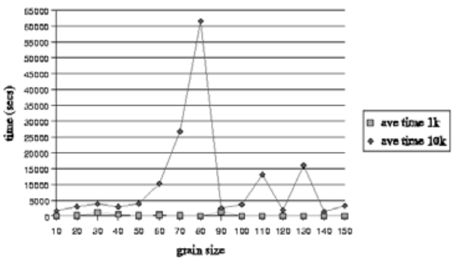

Figure 2: Optimization time: granulation and optimization ($C = 10^n$)

classification accuracy for each $C = 10^n$ (where $n = 1, 2, \ldots$), and they are shown in figures 2 and 1 respectively. Table 4 shows the performance comparison of SVM optimization time and classification accuracy for the cases with and without value granulation. All numbers plotted and appeared in those figures and the table are average numbers of ten trials.

We observe in this result that value granulation does not necessarily improve the learning result in $|D| = 1000$. This rather serves as an overhead to the SVM optimization. In the case of $|D| = 10000$, we observe some improvement (but not so significant), as a result of the value granulation, both in classification accuracy and optimization time.

$\|D\|$	Accuracy (%)	Training Time (sec)
1000	89:87	24:89
10000	75:83	13200:10200

Table 4: Performance Comparison (No Gr : $C = 60$)

4.2 Action Granulation vs. Information Granulation

We consider two action granulation in machine learning: ArraySVM and Granular Expert SVM (GESVM). ArraySVM consists of the three procedures:

1. *Stochastic (variable) granulation.* The data set D is divided into a set of granules G. Currently, we use random selection of records with uniform size (i.e. the uniform number of records per granule) and no duplication.

2. *SVM optimization for each granule.* As a result, there will be $|G|$ SVM classifiers.

3. *SVM classification by aggregation.*

$$h(\vec{x}) = h_i(\vec{x}),\ i = \arg\max_i [|h_i(\vec{x})|] \quad (2)$$

Similarly, GESVM consists of the following:

1. *Value granulation.* As described in section 2.2.

2. *Learning.*
   (a) If $|g \in G| = 0$, classifier $h_g(x)$ follows a general classification rule, e.g. $h_g(x) = h_{g'}(x)$ such that $\|g - g'\|$ is minimum.
   (b) If $|g \in G| \neq 0$ and the positive class is dominant in granule g, $h_g(x) = 1$. Likewise, $h_g(x) = -1$ if the negative class is dominant.
   (c) Perform SVM optimization if $|g \in G| \neq 0$ and no class is dominant in granule g.

3. *Classification* is performed by $h(x) = h_g(x)$ where $x \in g$.

Comparing the performance between information granulation and action granulation, we may observe more significant difference. Table 5 shows the performance comparison between GSVM (information granulation) and ArraySVM (action granulation). The classification accuracy of ArraySVM outperforms GSVM and it is consistent in the change of

$\|D\|$	Accuracy (%)	Training Time (sec)
1000	89:91	89:45
10000	83:91	10200:491

Table 5: Performance Comparison with ArraySVM (Information:Action)

$\|D\|$	Accuracy (%)	Training Time (sec)
1000	89:87	89:13
10000	83:93	10200:92

Table 6: Performance Comparison with GESVM (Information:Action)

data size from 1000 to 10000. The training time of ArraySVM is significantly advantageous, especially in the case of data size with 10000.

Table 6 shows the performance comparison between GSVM and GESVM. In this case, we observe a significant advantage on the training time. Though classification accuracy of GESVM does not always outperforms that of GSVM, it is indeed advantageous with the larger data size.

Overall, we see effectiveness of value granulation applied to SVM in general. Between two different application of value granulation to machine learning, action granulation clearly is more advantageous than information granulation in classification accuracy and, especially in training time.

4.3 Stochastic Granulation vs. Geometric Granulation

In ArraySVM and GESVM, we introduce two different granulation: stochastic (variable) granulation and geometric (value) granulation respectively. Stochastic granulation generates granules with uniform size through random selection of records from a data set. Geometric granulation, on the other hand, follows the procedure in section 2.2. The size of granules in geometric granulation varies.

Table 7 shows the performance comparison between ArraySVM (with stochastic granulation) and GESVM (with geometric granulation). First, action granulation appears to be more robust against the increase of data size

$\|D\|$	Accuracy (%)	Training Time (sec)
10000	91:92	491:92
250000	89:98	11900:1770

Table 7: Performance Comparison of Action Granulation (Stochastic:Geometric)

in comparison with information granulation. Second, geometric granulation (i.e. GESVM) appears to be more effective both in classification accuracy and optimization time. Among ten trials, GESVM has shown 99.7% of classification accuracy with the data size of 250000.

GESVM brings not only the significant advantages of optimization time due to its simplicity (i.e. domain segmentation) but also an advantageous classification accuracy, especially in a large scale case (i.e. 250000 records). As a consequence, we demonstrate that a geometric, value granulation (i.e. discretization and quantization) with an application of action granulation to machine learning is the most effective on both computation and classification accuracy, especially when handling a massively large data set.

4.4 A Simple Machine Learning Based on Value Granulation and Purity of Granules

SVM is known to be more powerful machine learning algorithm in comparison to others, due mainly to its generality. However, SVM optimization is rather computationally expensive and thus is less effective when handling a massively large data set. As being described in this paper, some additional framework is necessary in order to overcome the matter of massively large data set. As a consequence, it appears that the simplest granular computing approach such as value granulation is indeed effective.

We may pursue further possible effectiveness, that is the elimination of SVM optimization. SVM is known to be an instance-based machine learning algorithm such as k nearest neighbor algorithm (k-NN). Theoretically, SVM outperforms k-NN due to its generality. However, k-NN is still widely used and considered due to its simplicity and robustness against massively large data.

Here we propose a simple classification based on the purity of granules with respect to class frequencies. We call a granule pure if this consists of data that belongs to only one class. A classification is performed according to a class whise frequency is the highest within the granule. If the granule is empty, a general rule of classification, such as the one closest, is applied. Alternatively, this may consider unable to be classified depending on the problem domain and the system requirement. The learning is simply a matter of generating class histograms for granules and can thus be easily integrated with the value granulation described in section 2.2. This is very similar to k-NN except that the voting takes place within a granule instead of among k neighbors. However, its computational advantage in classification is significant because this does not have to search k nearest neighbors from samples. Instead, it simply looks into the class histogram of a granule. We formalize this machine learning as follows:

Name *Granular Classification by Majority (GrCM)*

Learning Generation of frequency histograms per granule $H_g(f(x))$ from each sample $< x, f(x) >$ during the value granulation described in section 2.2. (Note: $f(x)$ is a class label of x.)

Classification Given an arbitrary input x and its granule g such that $x \in g$, classification function $h(x)$ is specified as follows:

$$h(x) = \begin{cases} \arg\max_c[H_g(c)], & |g| > 0 \\ \text{ERROR or} & \\ \arg\max_c[H_{g'}(c)], & |g| = 0 \end{cases} \tag{3}$$

where $g' = \arg\min_{g'} \|g - g'\|$ for granule g.

Table 8 shows the performance comparison between value granulation with (GESVM) and without SVM (GrCM). It is somewhat disappointed, and expected at the same time, that GrCM does not necessarily overperform

$\|D\|$	Accuracy (%)	Training Time (sec)
10000	92:81	92:85
250000	98:84	1770:980

Table 8: Performance Comparison with no SVM (GESVM:GrCM)

GESVM. We, as a result, observe the power of generalization in SVM. On the other hand, We are still curious of how GrCM performs in larger data sets exceeding a million.

5 Intelligent System Development

Orthodox intelligent systems utilize IF-THEN rules in order to represent their intelligence. Those rules are extracted and constructed from a complex domain where rigorous analysis is not effective using machine learning algorithms. IF-THEN rules are rather descriptive than prescriptive and thus qualitative approaches are considered effective for their extraction and construction. Sugeno-Yasukawa's qualitative fuzzy modeling is indeed a case[9] in order to develop fuzzy logic systems. In this modelling, fuzzy sets for the IF-THEN rules need to be identified and a fuzzy clustering is used for that purpose. In Granular Computing, those fuzzy sets are considered as concept or variable granules.

We are now considering a decision support system for computer network management. This system supports decision making for computer networks in a timely manner. As previously described (e.g. the DARPA IDS evaluation data set), this domain is highly dimensional, i.e. complex, and usually generates massively large data. Because of this, most of qualitative modeling such as Sugeno-Yasukawa may not be effective with respect to its computation, e.g. fuzzy clustering.

We are currently developing a decision support system for computer network management as illustrated in figure 3. In order to handle this massively large and complex domain, value granulation is performed to generate IF-THEN rules. This result then is used in order to configure various network traffic classifiers such as tcpdump and snort. The visualizer visualizes states of computer networks

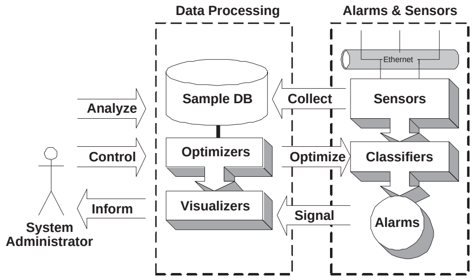

Figure 3: DSS for Network Management

based on alarms generated by the classifiers and the result of optimization (i.e. value granulation).

The highlight of this decision support system is the compatibility of information granule generated by value granulation and classifiers such as tcpdump and snort. Since the boundaries of this information granule are explicit value segments, they can be directly used as configuration of tcpdump and snort, that are essentially configured by means of crisp IF-THEN rules. Further, considering this system as an instance of GrCM, we can easily facilitate accountability of the configuration of tcpdump and snort based on the class histograms of GrCM.

6 Concluding Remarks

We discuss and demonstrate effectiveness of value granulation in machine learning. The effectiveness is essentially brought by both its low computation and size reduction. Although effectiveness of its size reduction is not always guaranteed, this study showed that there is a good chance of being effective.

Further, application of value granulation to machine learning is discussed. We find, by studying cases of SVM, that action granulation is more effective than information granulation. We also study a simple machine learning algorithm, GrCM, as an attempt of demonstrating significance of value granulation in machine learning (in comparison with that of SVM, i.e. generality). Unfortunately, this was not necessarily successful at this time. However, it is still hopeful that this may be demonstrated with much larger and complex data sets. At last, a decision support system is briefly introduced as a case study of GrCM.

There will be many future works, including more extensive analysis of size reduction, experiments of GrCM with much larger data sets and, most importantly, specification of a Granular Computing framework for massively large scale machine learning and data mining.

References

[1] A Bargiela and W Pedrycz. *Granular Computing: An Introduction.* Kluwer Academic, 2003.

[2] Thorsten Joachims. Svmlight web page.

[3] Richard Lippmann. The 1998 darpa/afrl off-line intrusion detection evaluation. In *First International Workshop on Recent Advances in Intrusion Detection*, 1998.

[4] J. Mill and A. Inoue. Reinforcement of a Support Vector Classifier. In *IPMU-2004*, 2004.

[5] J. Mill and A. Inoue. Support Vector Classifiers and Network Intrusion Detection. In *FUZZ/IEEE-2004*, 2004.

[6] J. Mill and A. Inoue. Granulation of machine learning. In *IPMU-2006*, 2006.

[7] J. Mill and A. Inoue. Constructing network intrusion alarms using granular computing and support vector machines. *IJUFKS*, to appear.

[8] C. V. Negoita and D. A. Ralescu. Representation theorems for fuzzy concepts. *Kybernetes*, 4(3):169–174, 1975.

[9] M. Sugeno and T. Yasukawa. A fuzzy logic based approach to qualitative modelling. *IEEE Transaction on Fuzzy Systems*, 1(1), 1993.

[10] L. A. Zadeh. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems*, 90, 1997.