 Impact of Parameter Tuning and Ensemble Methods on Fault Prediction Models

Extended Abstract†

|  |
| --- |
| Guru Darshan Pollepalli Manohara North Carolina State University, Raleigh USA gpollep@ncsu.edu |

Abstract

Two-dimensional[[1]](#footnote-2) arrays of bi-component structures made of cobalt and permalloy elliptical dots with thickness of 25 nm, length 1 m and width of 225 nm, have been prepared by a self-aligned shadow deposition technique. Brillouin light scattering has been exploited to study the frequency dependence of thermally excited magnetic eigenmodes on the intensity of the external magnetic field, applied along the easy axis of the elements.

***CCS Concepts*** •**Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; •**Networks** → Network reliability

***Keywords*** ACM proceedings, text tagging

ACM Reference format:

Anonymous Author(s). 1997. SIG Proceedings Paper in word Format. In *Proceedings of ACM Woodstock conference, El Paso, Texas USA, July 1997 (WOODSTOCK’97)*, 4 pages.

DOI: 10.1145/123\_4

1 Introduction

Ability for an organization to produce quality software products is highly correlated with their Quality Assurance methodologies. An important aspect of Software Quality team is to predict the chances of a bug based on heuristics and past knowledge. Software giants have shown increasingly more interest in creating models capable of encapsulating the previously known knowledge about the software in terms of predefined metrics and creating data models which can be used to predict the occurrence of faults in their future releases. Identifying faulty modules helps the QA team to redirect their efforts and resources. On the research front, majority of the studies performed in the field of Software Engineering encompass fault prediction and project planning with more number of papers published in the field of fault prediction. Menzies et al [9] suggest that the probability associated with the identification a bug using fault prediction models is relatively high (71%) when compared to traditional code review methodologies which result in a probability of around 60%. Code reviews increase the chances of human error rates as the code will have to be examined every line by line to look for errors. Hence using fault prediction models have significant advantage in terms of speed and accuracy when compared to traditional methods of code review.

A plethora of modelling techniques are available in a data scientist’s tool box to capture, analyze and predict different faults in a software system. Software metrics must be computed before the algorithm is run. Chidamber and Kemerer [5] provide a suite of metrics as shown in Figure 1 for object oriented paradigms which can be used to evaluate the fault proneness of a class. Computing the CK metrics would involve examining the bug repositories like Bugzilla or GitHub and extracting the classes responsible for the bugs to compute the number of bugs associated with a class. Tibor et al [10] examined Mozilla repository where they used Bugzilla to obtain bug reports of projects associated with Mozilla and manually examined nearly 3000 bug reports to extract the classes responsible for the bugs. Many famous open-source projects publish CK metrics online to be used by data scientists to train their models. This project uses PROMISE data repository and Seacraft repository to obtain CK metrics for many open-source projects like log4j,

2 CRITERIA

2.1 Readability

Readability of a data model is highly instrumental in many disciplines of Software Engineering and its related focuses. Project management tasks heavily rely on the organization and transparency of the model generated by the algorithms to effectively allocate resources and make minor adjustments to the model in order to provide the necessary flexibility for accomplishing the organizational goals. Many project management softwares are equipped with capturing microscopic details about products which are then translated into usable data models where managers can transform the data inputs into strategic goals and perform informed decisions. Project managers are responsible for overseeing the project milestones and deliverables. They are also responsible to dynamically adjust the resources based on availability without having request the complete recompilation of the data model. Data models which provide little to none readability can slow down the process significantly. In many open-source software development cycles where developers are available on a voluntary basis and collaboration from developers becomes an essential component of project management, the need to have a readable and interpretable model becomes a necessity than a cosmetic. In cases where a developer is unavailable, the manager can quickly consult the data model to reallocate the resources on demand. Human readable models like Decision Trees provide better insights into the tasks than models like Artificial Neural Networks. Graphical illustration provided by Decision Trees increase their ease of use and help mangers to analyze the models better. Readable models can also be used to methodologically scrutinize the various inter dependencies between projects and help them create varied levels of focus for individual projects.

2.2  Actionable Conclusions

Sometimes it is imperative to know the differences between the classifications than the classification themselves. That is, we can benefit much more by knowing the difference (in attributes) which caused a data point to be classified into one of the categories (say X) and not another category (say Y). In bug prediction algorithms, contrasts between Bug and No-Bug categories can be examined by building a contrast learner. For example, in a bug prediction algorithm using Chidamber & Kemerer metrics [5], one can look at the classes which have bugs and say that a high value in CBO (Coupling Between Object classes - which can be qualitatively defined as how much a method in a class is dependent on methods or variables in another class) tends to increase the probability for a class to be bug prone [6]. In case of project management, one can assess if the number of hours spent by a developer on a task has a direct impact on his code being bug free or bug prone. Contrast learners can be programmed further to recognize the importance of the classification categories and generate contrasts with respect to only those categories. This can highly reduce computational costs associated with the learner as less data is examined to provide a conclusion.

2.3  Learnability and Repeatability of the Results

TODO

2.4 Multi-goal reasoning

Virtually every real-world problem is associated with multiple goals. Many decision-making problems in the business world are formulated as multi-goal reasoning problems. In Economics and Finance world, the determination of policies and portfolios involve attaining a delicate balance between multiple (often conflicting) goals and the variant of multi-goal reasoning algorithms used play a very important role in determining the outcomes associated with the problem. Multiple iterations of the model are simulated to rigorously test and evaluate the performance of the model. Given that more than one solution is possible in multi-goal reasoning problems, the user is required to find tradeoff between the solutions and pick the solution best suited for the use case. Repeated application of the algorithm is required to obtain the multiple solutions. One of the methods of obtaining solutions for multiple goal reasoning is by scalarizing multiple goals into single goals [3] (such as denomination scores) and then use the algorithms designed for single goal reasoning to obtain solutions for the multi-goal problem at hand. Evolutionary algorithms can be used in place which provide many unique advantages including efficient computation and better convergence of solutions [3].

2.5 Anomaly Detection

Anomaly detection can be defined as a set of paradigms capable of detecting any uncommon data which do not conform to the expected behaviour of the data (outliers). Anomaly detection has several applications in medical sciences where detection of tumours can be performed using anomaly detection MRI scans, credit card fraud can be detected if there is a huge change the expenditures and very often it has been used to detect the health of component machine of a control system by continuously observing its internal parameters. The anomalies can range from context-specific in nature to collective anomalies which can contain a stream of data points uncommon to the data model. Many safeguards have been designed in response to anomaly detection like blocking a credit card till the user can verify the exorbitant purchase or alerting the operator before a control system fails. Contextual anomaly detection can be used in case the anomaly detected must be examined in its local environment like a sudden increase in network bandwidth usage during sports broadcasts should not be cause any flags to be raised even though the bandwidth usage is above the normal limits.

2.6 Incremental Learning

In today’s world, the database sizes increase at an alarming rate. Many companies are heavily investing in technologies capable of scaling their data learning algorithms to adapt to massive amounts of incoming training data [1]. Extensive research is done to find frameworks for existing algorithms to augment the ability to learn incrementally and provide continuous availability of data. Reconstruction of Data models upon receipt of new training data can be done either by building the model from scratch or appending the new data to the data model without sacrificing the existing knowledge base. The former method becomes computationally expensive when there are massive amounts of data involved (like Big Data). Hence the data model must be incrementally trained to update the knowledge base over time to cater to the real-time requirements of the task at hand. Alternatives include periodic retraining using batches of data. A crude comparison can be made to biological intelligent systems where the information is learnt over their lifetime to accumulate behaviors and develop associations to construct goal oriented behavior [2]. There are situations where the data used to train the data model may not be available for retraining the data. In such cases, incremental learning comes to the rescue.

2.7 Shareable

Data privacy is one of the very key factors to be considered while using data mining algorithms. The data used to train the data can be private to an organization but the data will be requested by a third party for creating data models. Essential protocols will have to be developed for sharing the data and must capture rules governing how the data is shared and how much of the data is shared. The protocol must attain a delicate balance between the reduction of dispensable data and concentrate on the vital data points which adequately describe the data model. Data obfuscation techniques will have to be used to mask the actual data but still maintaining the required parts to construct the data model. Increasing size of databases and data sets pose a further problem of how much data must be shared to create an efficient data model.

2.8 Context Aware

A gamut of today’s state-of-the art data learning models visualize the data as an aggregated information and pay little to no heed to the sub-portions of the data which were obtained under different environments (contexts). Viewing the data set as being constituted of homogenous populations and deriving conclusions based on the assumptions can often lead to critical errors in judgements leading to the depletion of accuracy of the data model. As an example, consider the following statement “Exercise and athletics reduces the risk of diabetes”. The statement generalizes by saying that exercise and athletics is always good. This might be true in case of a general population but the statement must be re-evaluated for sub-populations which contains patients recently treated for heart attack or people who had recently undergone surgery. The differences these sub-populations exhibit can be encoded into a context which specifies the type of population under observation. Even though these sub-populations form a smaller part of the overall populations, it is imperative that the data models are aware of the contexts and accordingly tune the control parameters of the data model or generate different data models based on the contexts. Since the exact number of context is unknown beforehand, it is hard to determine the partitions to capture all the contexts. Algorithms like Mini Batch K-Means can be used to cluster the data and then apply the algorithm on each individual cluster to create different data models aware of the local environments.

2.9 Self-tuning

Attributes of a model like its accuracy, performance, etc. are highly reliant on the values of the control parameters chosen for the algorithm at the beginning of the data modelling. In case of KNN, the accuracy of the data model is determined by the value of K chosen before the algorithm is started. For very low values of K, the algorithm tries to overfit to the training data and the effect of noise is predominantly amplified. For large values of K, the algorithm tries to underfit the model but irons out the noisy data points. In order to obtain the combination of parameters which provides the best score, the algorithm will be required to run with different parameter settings or run all possible combinations of the parameters which becomes computationally intensive [4] (but provides an exhaustive search of all the possible combinations which can be used to optimize the performance of the model). Evolutionary algorithms like Differential Evolution decrease the convergence time by orders of magnitude using mutation and cross-overs of candidate solutions. Tuning has a direct impact on the conclusion obtained using the learners. Well-tuned learners tend to provide better accuracy and precision compared to other learners and the results of many research papers can be directly challenged by testing with tuned data models [4].

3 Results and Discussion

3.1 Magnetization Curves and MFM Characterization

The major hysteresis loop measured by MOKE, plotted in [Fig. 1](#fig1), displays a two-step switching process due to the distinct magnetization reversal of the Py and Co sub-elements, characterized by a different coercivity. As the field is reduced from positive saturation (upper branch of the M-H loop), a 100% remanence is attained. On reversing the applied field, one observes a drop of the magnetization (between −240 and −370 Oe), proportional to the Py magnetization fraction within each bi-component unit (about 36%) in good agreement with experimental result (about 40%).

To directly visualize the evolution of the magnetization in the Py and Co subunits of our bi-component dots during the reversal process, we performed a field-dependent MFM analysis whose main results are reported in [Fig. 2](#fig2). At large positive field (*H*= +800 Oe, not shown here) and at remanence (** point of the hysteresis loop of [Fig. 1](#fig1)), the structures are characterized by a strong dipolar contrast due to the stray fields emanated from both the Py and Co dots.



**Figure 2.** MFM images of the bi-component Py/Co dots for different values of the applied magnetic field which are indicated by greek letters along both the major and minor hysteresis loop.

This is a clear indication that both the Py and Co sub-elements are in a single domain state where Py and Co magnetizations are all oriented with their magnetic moment along the chain and field direction. At point ** (*H*= −372 Oe) of the hysteresis loop, where the plateau is observed in the *M*-*H* loop, the dark and bright spots of the Py dots are reversed with respect to those of Co, accounting for an antiparallel relative alignment of magnetization.

At relatively large negative fields (point ** *H*= −770 Oe) the magnetization reversal is completed and the magnetization of the two adjacent sub-elements are saturated in the negative direction. The ground state remains unchanged when the field is now reduced to zero, i.e. remanent state coming from negative saturation, as confirmed by the MFM image taken at point  of [Fig. 1.](#fig1)

We have also used MFM to measure the magnetic configurations along the minor hysteresis loop, described above. Once the AP ground state has been generated at *H*= −500 Oe, the applied field is increased in the positive direction. The MFM image taken at point **' of [Fig. 2](#fig2), remains unchanged until the magnetic field is increased up to +300 Oe where the Py magnetization reverses its orientation and returns to be aligned with that of Co dots. On the basis of the above MFM investigation, one can say that the structures are always in a single domain state, while the relative magnetization orientation between the adjacent Py and Co elements depends on both the field value and the sample history.



**Figure 3.** Dependence of the magnetic eigeinmode wave frequency on the applied field strength.



**Figure 4.** Calculated spatial distribution of the in-plane dynamic magnetization.

3.2 Field Dependent BLS Measurements and DMM Calculations

[Fig. 3](#fig3) displays the frequencies of BLS peaks plotted as a function of the applied field magnitude starting from positive values. The field is then decreased and reversed following the upper branch of the hysteresis loop, shown in the same figure. Up to five peaks are measured in the spectra, as shown in spectrum measured at H = 0 Oe in the [Fig. 3](#fig3) inset, and their field evolution analyzed over the whole field range investigated. The detected modes are identified and labeled on the basis of their calculated spatial profiles, shown in [Fig. 4](#fig4) for *H*= 500 and −500 Oe. They exhibit marked localization into either the Co or the Py dots, as stated at the end of the previous Section, were it was introduced the labelling notation containing the dominant localization region (either Py or Co) and the spatial symmetry (EM, F, DE, etc).

When the dots are in the P state, up to five modes were detected in BLS spectra. On the basis of the calculated profiles (right panel of [Fig. 4](#fig4)), we identified in the P state the two modes at lowest frequencies as the EM(Py) and the F(Py), with a very small spin precession amplitude into the Co dot. This is because for this material we are below the frequency threshold for the existence of spin waves. A similar effect has been observed in periodic array of alternating Permalloy and Co nanostripes

Note that the nodal lines present in the spatial profile of the F (Co) mode perpendicular to the long axis of the ellipse do not correspond to a real change of sign of the dynamic magnetization and are due to the partial hybridization of the F mode with higher-order modes having frequencies close to the one of the F mode. Interestingly, the frequency slope of modes localized into the Co dots is larger than that of Py modes, due to larger values of the Co magnetization and gyromagnetic ratio. An overall good agreement between the calculated (dotted curves) and measured frequency (full points) has been achieved (see [Fig. 3](#fig3)) even if some discrepancies are observed for the frequency of the EM and 1DE (Py) modes.

The corresponding spatial profiles of the modes are shown in the left panels of [Fig. 4](#fig4). Here one can see that the only mode which is purely localized in one dot is the EM of Co, because now it is sub-threshold for Py. A further reduction of *H*, which is sufficient to cause the Co magnetization reversal, produces a P state at negative fields and the frequency starts to increase again as a function of the applied field. In this field range the frequencies of modes in the Py dots monotonously increase in a way similar to that measured in the P state for positive field values while an abrupt change in the frequency of Co modes occurs.

Notice that if one stops increasing the negative field to about −300 Oe and comes back towards positive applied fields, BLS measurements can be performed following the minor hysteresis loop. This method permits to study, for example, the magnetization dynamics at remanence (without any external applied magnetic field) when the system is in the AP state (see MFM image ' in [Fig. 2](#fig2)), a configuration which cannot be achieved at remanence along the major M-H loop. In [Fig. 5](#fig5) we show the modes frequency measured along the minor loop (full points) and compare them with values measured along the major M-H loop (open points).

By inspection of the frequency slope of the modes, one can immediately understand the localization of modes into dots of different materials looking at their slope.

In particular, for three (two) modes we measure a negative (positive) frequency slope with an almost linear dependence on *H*. It is evident that modes with negative frequency slope are modes localized into the Py dot (EM, F and 1DE) while the two with positive slope are the F(Co) and the EM(Co) modes.

3.3 Analysis of the Dynamic Coupling as a Function of the Gap Size

One interesting point which emerges from analysis of [Figs. 3](#fig3) and [4](#fig4) is that the frequency values of the eigenmodes are not the same at +500 Oe and at −500 Oe. This is expected for modes localized into the Co elements, since the external field is either parallel or antiparallel to their magnetization. However, for those mode localized into the Py sub-element, one could have predicted to find the same frequency values at ±500 Oe, unless the dipolar coupling arising from the adjacent Co dot plays a significant role. In fact, as seen in [Figs. 3](#fig3) and [4](#fig4), reversing the field from +500 to −500 Oe, the frequencies of EM(Py) and 1DE(Py) modes increase by about 0.2 GHz and 0.6 GHz, respectively, while that of F(Py) decreases by 0.25 GHz. The reason of this complex behavior will be addressed in the following, analyzing the interplay of both static and dynamic dipolar coupling between the adjacent Py and Co dots [Table 2](#tb2). Notice that if one stops increasing the negative field to about −300 Oe and comes back towards positive applied fields, BLS measurements can be performed following the minor hysteresis loop.



**Figure 5.** Full point are the frequencies measured along the minor hysteresis.

**Table 2.** Comparison of Coefficients from Atomistic

|  |  |  |
| --- | --- | --- |
| Atm | MS-CG | MS-CG/DPD |
| 1.78 | 14.32 | 1.74 (−2%) |
| 0.43 | 31.00 | 0.40 (−7%) |
| 0.062 | 15.61 | 0.048 (−23%) |
| 0.032 | 9.76 | 0.024 (−24%) |
| 0.020 | 4.66 | 0.015 (−25%) |
| 0.012 | 2.32 | -"- |
| 0.0076 | 0.016 | -"- |



**Figure 6.** Calculated frequency evolution of modes detected in the BLS spectra.

In [Fig. 6](#fig6) the calculated frequencies of the most representative eigenmodes at +500 Oe (FM state) and – 500 Oe (AP state) are plotted as a function of the gap size d between the Py and Co sub units (please remind that in the real sample studied here, d = 35 nm). As a general comment, it can be seen that the frequencies for the system in the AP state are more sensitive to d than those of the P state. In particular, the lowest three frequency modes of the AP state (EM(Co), EM(Py) and F(Py)) are downshifted with respect to the case of isolated elements (dotted lines) and show a marked decrease with reducing d, while the two modes at higher frequencies (F(Co) and 1DE(Py)) have an opposite behavior even though they exhibit a reduced amplitude. In the P state (right panel), the modes concentrated into the Py dots exhibit a moderate decrease with reducing d, while an opposite but less pronounced behavior is exhibited by the F(Co) mode.

1. Never, ever use vertical rules.
2. Never use double rules.

4 Conclusions

In summary, we have performed both an experimental and theoretical study of the spin eigenmodes in dipolarly coupled bi-component cobalt and permalloy elliptical nanodots. Several eigenmodes have been identified and their frequency evolution as a function of the intensity of the applied magnetic field has been measured by Brillouin light scattering technique, encompassing the ground states where the cobalt and permalloy dots magnetizations are parallel or anti-parallel, respectively. In correspondence to the transition between the two different ground states, the mode frequency undergoes an abrupt variation and more than that, in the anti-parallelstate, the frequency is insensitive to the applied field strength. The experimental results have been successfully interpreted by the dynamic matrix method which permits to calculate both the mode frequencies and the spatial profiles.

Acknowledgments

This work was partially supported by the MIUR-PRIN 2010–11 Project 2010ECA8P3 “DyNanoMag” and by the National Research Foundation, Prime Minister's office, Singapore under its Competitive Research Programme (CRP Award No. NRF-CRP 10-2012-03).

A Headings in Appendices

The rules about hierarchical headings discussed above for the body of the article are di.erent in the appendices. In the appendix environment, the command section is used to indicate the start of each Appendix, with alphabetic order designation (i.e., the first is A, the second B, etc.) and a title (if you include one). So, if you need hierarchical structure within an Appendix, start with subsection as the highest level. Here is an outline of the body of this document in Appendix-appropriate form:

A.1 Introduction

A.2 Experimental And Computational Details

A.2.1 Sample Fabrication

A.2.2 Quasi-Static Measurements: MOKE and MFM

Component Structures

Magnetization

A.2.3 Dynamic Measurements: BLS

A.2.4 Ground-State Magnetization Determination and DMM Micromagnetic Simulations

Determined

Micromagnetic

A.3 Results And Discussion

A.3.1 Magnetization Curves and MFM Characterization

A.3.2 Field Dependent BLS Measurements and DMM Calculations

A.3.3 Analysis of the Dynamic Coupling as a Function of the Gap Size

A.4 Conclusions

A.5 Acknowledgments

A.6 References

References

|  |  |
| --- | --- |
| [1] | Patricia S. Abril and Robert Plant. 2007. The patent holder’s dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan. 2007), 36–44. DOI: http://dx.doi.org/10.1145/1188913.1188915 |
| [2] | I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. 2002. Wireless Sensor Networks: A Survey. *Comm. ACM* 38, 4 (2002), 393–422. |
| [3] | David A. Anisi. 2003. *Optimal Motion Control of a Ground Vehicle*. Master’s thesis. Royal Institute of Technology (KTH), Stockholm, Sweden. |
| [4] | P. Bahl, R. Chancre, and J. Dungeon. 2004. SSCH: Slotted Seeded Channel Hopping for Capacity Improvement in IEEE 802.11 Ad-Hoc Wireless Networks. In *Proceeding of the 10th International Conference on Mobile Computing and Networking* (MobiCom’04). ACM, New York, NY, 112–117. |
| [5] | Kenneth L. Clarkson. 1985. *Algorithms for Closest-Point Problems (Computational Geometry)*. Ph.D. Dissertation. Stanford University, Palo Alto, CA. UMI Order Number: AAT 8506171. |
| [6] | Jacques Cohen (Ed.). 1996. Special Issue: Digital Libraries. *Commun. ACM* 39, 11 (Nov. 1996). |
| [7] | Bruce P. Douglass. 1998. Statecarts in use: structured analysis and object-orientation. In *Lectures on Embedded Systems*, Grzegorz Rozenberg and Frits W. Vaandrager (Eds.). Lecture Notes in Computer Science, Vol. 1494. Springer-Verlag, London, 368–394. DOI: http://dx.doi.org/10.1007/3-540-65193-429 |
| [8] | Ian Editor (Ed.). 2008. *The title of book two* (2nd. ed.). University of Chicago Press, Chicago, Chapter 100. DOI: http://dx.doi.org/10.1007/3-540-09237-4 |

1. \*Produces the permission block, and copyright information

   †The full version of the author’s guide is available as acmart.pdf document

   It is a datatype.

   Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

   WOODSTOCK’97, July 2016, El Paso, Texas USA

   © 2016 Copyright held by the owner/author(s). 123-4567-24-567/08/06. . . $15.00

   DOI: 10.1145/123\_4 [↑](#footnote-ref-2)