Buletin Teknologi MARDI Bil. 16 (2019) Khas Ternakan Lestari: 125 – 134

Kaedah, potensi dan cabaran analisis data besar dalam sains ternakan

(Techniques, potentials and challanges on big data analysis in livestock science)

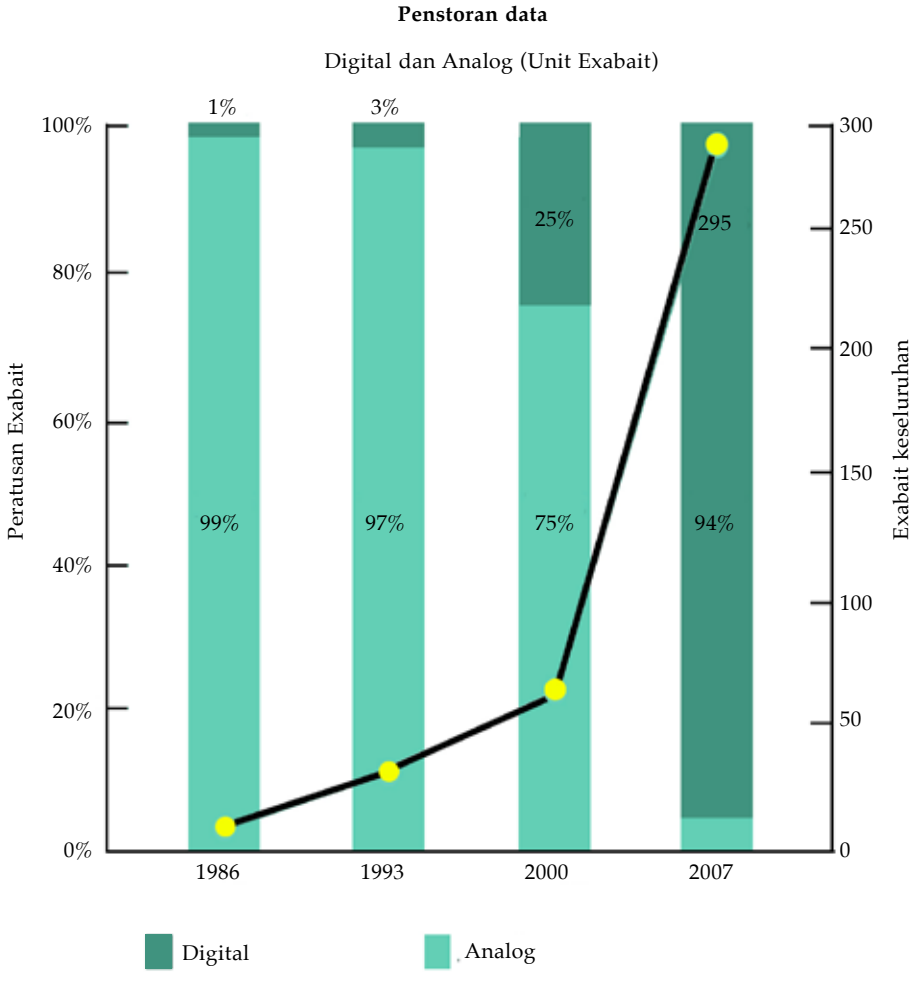
Mohd Azri Azman

Pengenalan

Terma data besar (*big data*) sering diguna pakai dalam kalangan saintis data dan pakar teknologi maklumat dalam menerangkan kepentingan pengurusan dan analisis data yang besar dan banyak. Terma ini dipopularkan selari dengan lambakan data secara eksponen, pengurangan kos teknologi yang digunakan untuk mengumpul dan memproses data serta kesedaran yang timbul mengenai potensi data-data yang besar dalam membuat keputusan. Penggunaan terma ini selari dengan konsep pertanian digital dan pertanian pintar, yang mana penggunaan internet benda [*Internet of Things* (IoT)] bersama-sama perkakasan seperti sensor dan datalog menjana jutaan data-data mentah yang memerlukan analisis dan pemprosesan untuk diterjemahkan dalam bentuk yang mudah difahami.

Maksud data besar

Data besar merujuk secara spesifik kepada satu koleksi set data yang sangat besar sehinggakan saiz data tersebut tidak dapat ditampung oleh komputer desktop ataupun komputer riba dan memerlukan keupayaan kerangka utama (*mainframe*) atau pengkomputeran awan (*cloud computing*) untuk diproses. Data dalam lajur (p) dan baris (n) adalah sangat besar dan mengehadkan pemerhatian secara visual. Selain itu, ciri-ciri data besar selalunya adalah data mentah yang belum diproses atau menjalani proses normalisasi serta mengandungi data hingar dan data pendua. Justeru, pengeditan data besar diperlukan sebelum sesuatu model disuaikan ke atas data tersebut. Oleh kerana takrifan 'besar' dalam konteks data analisis ini bergantung kepada keupayaan pengiraan sesuatu data, maka data besar boleh ditakrifkan sebagai data yang mengambil ruang melebihi sepertiga daripada memori capaian rawak [*random-access memory* (RAM)] satu komputer semasa membuat analisis tersebut. Selain itu, walaupun pemvisualan data memainkan peranan penting dalam membuat kesimpulan dan mengenal pasti ciri-ciri data, saiz data itu sendiri mengehadkan keupayaan memplotkan data tersebut untuk memberi gambaran besar kepada pengguna. Salah satu cara untuk memahami data besar ini dan mentransformasikan



Rajah 1. Perkembangan pertumbuhan data digital

data besar ini kepada ilmu pengetahuan yang dapat difahami adalah dengan menggunakan kaedah analisis yang sesuai dengan ciri-ciri data besar yang diperoleh.

Dimensi data besar (3V)

Data besar boleh ditakrifkan dengan menggunakan konsep 3V iaitu jumlah (*volume*), kepantasan (*velocity*) dan kepelbagaian (*variety*).

- Jumlah: Jumlah data bertambah secara eksponen. Data besar merujuk kepada dataset yang melebihi kemampuan kebanyakan perisian pangkalan data tipikal. Definisi ini adalah subjektif bergantung kepada kegunaan data tersebut dan aplikasinya pada industri. Sebagai contoh, data-data '-omics' yang bertambah dengan pesat yang terhasil daripada kaedah penjujukan DNA generasi baharu yang semakin pantas.

- Kepantasan: Menerangkan tentang kadar perubahan data yang mana jumlah data berubah secara dinamik. Dimensi ini merujuk kepada keupayaan memahami dan bertindak balas terhadap satu peristiwa semasa ia berlaku.
- Kepelbagaian: Data besar wujud dalam pelbagai bentuk serta format dan wujud dalam bentuk berstruktur dan tidak berstruktur. Kepelbagaian format data meningkatkan lagi kekompleksan data dari segi penstoran dan analisis. Contoh data berstruktur ialah data yang disimpan dalam Sistem Pangkalan Data Berkait [*Relational Database Management System* (*RDMS*)] dan data daripada lembaran sebaran (*spreadsheet*), manakala data tidak berstruktur adalah seperti imej, video dan dokumen teks.

Kaedah analisis data besar

Analisis data besar merangkumi gabungan pelbagai disiplin ilmu, antaranya statistik, perlombongan data, pembelajaran mesin, rangkaian neural, analisis rangkaian sosial, pemprosesan signal, pengecaman corak, kaedah pengoptimuman dan pendekatan visual. Terdapat beberapa kaedah spesifik dalam disiplin yang disebutkan dan bertindih antara satu sama lain. Berikut adalah beberapa kaedah penting dalam analisis data besar yang sering digunakan:

Kaedah pengoptimuman

Kaedah pengoptimuman digunakan untuk menyelesaikan masalah kuantitatif dalam pelbagai bidang seperti fizik, biologi, kejuruteraan dan ekonomi. Pengoptimuman stokastik seperti pengaturcaraan genetik dan pengaturcaraan evolusi digunakan untuk menyelesaikan masalah berkaitan alam semula jadi. Namun, penggunaan analisis ini membabitkan penggunaan memori komputer yang tinggi dan kompleks serta memakan masa yang banyak. Penyelesaian bagi masalah ini dilaksanakan dengan aplikasi algoritma koperatif dan penggunaan pengkomputeran gabungan seperti pengkomputeran awan dan pengkomputeran grid.

Statistik

Statistik ialah sains untuk mengumpul, mengurus dan menginterpretasi data. Kaedah statistik digunakan untuk mengeksploitasi hubungan korelasi dan kausal antara objektif berbeza. Teknik statistik standard tidak sesuai untuk menguruskan data besar, akan tetapi para penyelidik telah mencadangkan lanjutan standard statistik atau kaedah statistik yang baharu. Contoh cadangan yang telah dikemukakan dan digunakan adalah algoritma anggaran yang efisien untuk regresi monotonik multivariasi berskala besar yang merupakan pendekatan untuk menganggarkan fungsi-fungsi yang monotonik berkenaan dengan pemboleh ubah input. Selain itu,

kaedah analisis lain adalah analisis statistik berasaskan data memberi tumpuan kepada skala dan pelaksanaan algoritma statistik secara selari.

Perlombongan data

Perlombongan data adalah satu kaedah pengekstrakan maklumat yang penting daripada data. Teknik-teknik perlombongan data antara lainnya adalah analisis kluster, klasifikasi, regresi dan kaedah pembelajaran mesin menggunakan teknik peraturan perhubungan (*association rule learning*). Perlombongan data besar lebih mencabar berbanding dengan perlombongan data tradisional, sebagai contoh; cara semula jadi pengklusteran data besar adalah dengan memperluaskan kaedah sedia ada (seperti pengklusteran secara hierarki, *K-Mean* dan *Fuzzy CMean*) supaya mereka dapat mengatasi masalah beban kerja yang besar. Dalam disiplin bioinformatik, penjanaan data biologi secara eksponen mendorong kepada perubahan paradigma daripada pendekatan kajian gen tunggal kepada pendekatan yang menggabungkan analisis pangkalan data integratif dan perlombongan data. Paradigma baharu ini membolehkan kajian sintesis berskala besar fungsi genom dapat dilakukan.

Pembelajaran mesin

Manusia mampu membuat hubung kait corak dan hubungan dengan data, tetapi tidak dapat memproses sejumlah besar data dengan cepat. Mesin sebaliknya sangat mahir dalam memproses sejumlah besar data dengan cepat, tetapi hanya jika mereka tahu bagaimana untuk melakukannya. Jika pengetahuan manusia dapat digabungkan dengan kelajuan pemprosesan mesin, ia akan dapat memproses sejumlah besar data tanpa memerlukan campur tangan daripada manusia. Ini adalah konsep asas pembelajaran mesin.

Ciri yang paling penting dalam pembelajaran mesin adalah kebolehan membuat keputusan pintar secara automatik. Berikut adalah kaedah bagaimana pembelajaran mesin dilakukan dalam memproses data besar.

1) Klasifikasi (pembelajaran mesin diselia)
   Klasifikasi adalah teknik pembelajaran mesin yang diselia yang mana data diklasifikasikan ke dalam kategori relevan yang terdiri daripada dua langkah:

   a) Sistem disuap dengan data-data latihan yang telah dikategorikan atau dilabelkan, supaya dapat memahami kategori yang berlainan.

   b) Sistem disuap dengan data yang tidak dikenali oleh sistem, tetapi data yang sama untuk klasifikasi dan berdasarkan pemahaman yang dibangunkan daripada

data-data latihan dalam langkah (a) sebelum ini, algoritma sistem ini akan mengklasifikasikan data tersebut.

Contoh aplikasi kaedah ini adalah seperti dalam sistem penapisan spam e-mel. Ambil perhatian bahawa klasifikasi boleh dilakukan untuk dua atau lebih kategori. Dalam proses klasifikasi yang mudah, mesin akan disuap dengan data berlabel semasa latihan dan mesin akan belajar memahami klasifikasi data berdasarkan data yang diberi. Mesin itu kemudian akan disuap dengan data tidak berlabel, yang mana sistem akan mengklasifikasikan data tersebut secara sendirinya berdasarkan latihan yang telah diterima sebelumnya dengan data berlabel.

2) Pengklusteran (pembelajaran mesin tidak diselia)
Pengklusteran adalah teknik pembelajaran tidak diselia yang mana data dibahagikan kepada kumpulan yang berlainan supaya data dalam setiap kumpulan mempunyai sifat yang sama. Tiada langkah pembelajaran terdahulu diperlukan oleh sistem. Dalam kaedah ini, kategori dihasilkan berdasarkan pengelompokan data. Bagaimana data dikumpulkan bergantung pada jenis algoritma yang digunakan. Setiap algoritma menggunakan teknik yang berbeza untuk mengenal pasti kluster. Pengklusteran biasanya digunakan dalam perlombongan data untuk mendapatkan pemahaman tentang sifat-sifat suatu dataset yang diberikan. Selepas sistem memahami sifat dataset ini, klasifikasi boleh digunakan untuk membuat ramalan yang lebih baik mengenai data yang serupa, tetapi baharu atau data yang tidak jelas.

3) Pengesanan *outlier*
Kaedah ini adalah proses untuk mencari data yang jauh berbeza atau tidak konsisten dengan data lain dalam dataset yang diberikan. Teknik pembelajaran mesin ini digunakan untuk mengenal pasti anomali, keabnormalan dan data yang jauh terasing daripada plot data-data lain. Pengesanan *outlier* adalah berkait rapat dengan kaedah klasifikasi dan pengklusteran, walaupun algoritmanya hanya tertumpu pada mencari nilai yang tidak normal. Aplikasi untuk pengesanan *outlier* termasuk diagnosis perubatan, analisis data rangkaian dan analisis data sensor.

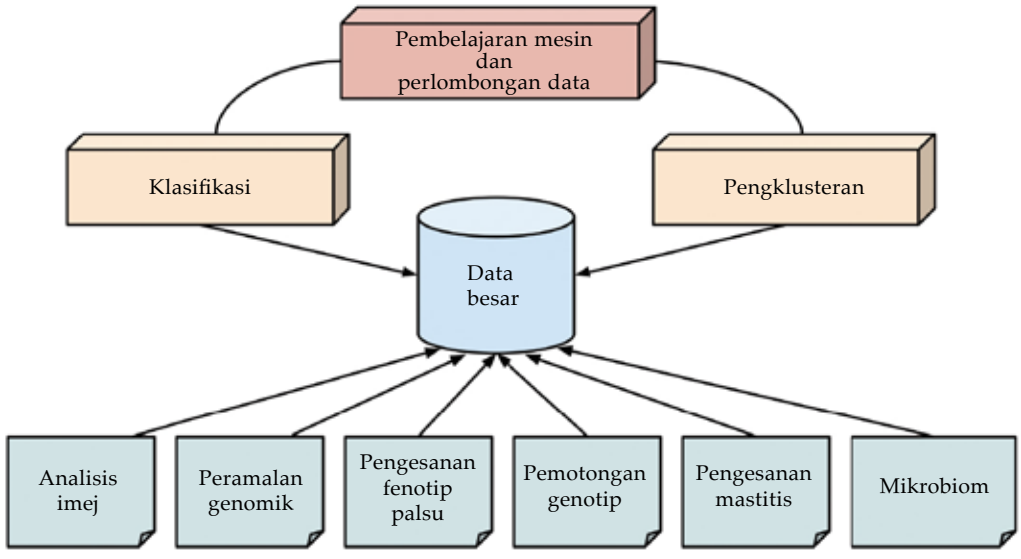
4) Penapisan
Penapisan adalah proses automatik mencari item yang relevan daripada sekumpulan item yang boleh ditapis sama ada berdasarkan tingkah laku subjek atau dengan menyesuaikan kelakuan pelbagai subjek. Penapisan biasanya digunakan melalui dua pendekatan seperti yang berikut:
- penapisan kolaboratif
- penapisan berasaskan kandungan

Penapisan kolaboratif adalah teknik penyaringan item berdasarkan kolaborasi atau penggabungan tingkah laku subjek sasaran dengan subjek yang lain. Penapisan kolaboratif hanya berdasarkan kesamaan antara tingkah laku subjek sahaja. Ia memerlukan banyak data tingkah laku subjek untuk menapis item dengan tepat.

Penapisan berdasarkan kandungan adalah teknik penapisan item yang memberi tumpuan kepada persamaan antara subjek dan item. Profil subjek dibuat berdasarkan tingkah laku subjek masa lalu. Bertentangan dengan penapisan kolaborasi, penapisan berasaskan kandungan hanya dibuat untuk pilihan subjek individu dan tidak memerlukan data tentang subjek lain. Sistem pengesyoran meramalkan pilihan subjek dan menghasilkan cadangan untuk subjek dengan sewajarnya. Sistem pengesyoran biasanya menggunakan penapisan kolaborasi atau penapisan berdasarkan kandungan untuk menghasilkan cadangan. Ia juga mungkin berdasarkan hibrid penapisan kolaborasi dan penapisan berasaskan kandungan untuk menyempurnakan ketepatan dan keberkesanan cadangan yang dihasilkan.

5) Rangkaian neural
Rangkaian neural adalah satu set algoritma, dimodelkan secara kasar mengikut otak manusia, yang direka bentuk untuk mengenali corak. Rangkaian neural menginterpretasi data sensori melalui pembelajaran mesin, pelabelan atau pengklusteran data input mentah. Corak yang mereka kenali adalah bernombor dan disimpan dalam bentuk vektor yang mana semua data, sama ada imej, bunyi, teks atau masa mesti diterjemahkan.

Potensi analisis data besar dalam sains ternakan
Ramalan genomik
Disiplin genetik adalah antara disiplin ilmu yang terawal mengguna pakai kaedah pembelajaran mesin dan perlombongan data dalam analisis data besar, dalam konteks peramalan fenotip. Antaranya penggunaan kaedah pembelajaran mesin dalam prosedur klasifikasi memilih polimorfisme nukleotida tunggal [*single nucleotide polymorphism*



Rajah 2. Gambaran keseluruhan analisis data besar dalam sains ternakan

(SNP)] dalam pemilihan genomik ternakan seperti yang telah dilaporkan dalam jurnal antarabangsa. Kaedah pembelajaran mesin memudahkan saintis haiwan berurusan dengan data heterogen yang banyak dan telah dikumpul selama bertahun-tahun. Kaedah pembelajaran mesin boleh digunakan untuk menghasilkan ramalan genetik yang baik, mencari hubungan yang tidak diketahui di kalangan fenotip atau antara fenotip dan pemboleh ubah persekitaran dan memantau kelebihan dan kekurangan sesuatu fenotip antara baka ternakan tertentu.

Analisis imej

Berat badan ternakan sangat penting untuk pengurusan pemakanan dan pembiakan kerana ia merupakan petunjuk langsung pertumbuhan haiwan, status kesihatan dan kesediaan untuk pasaran. Oleh itu, anggaran berat badan yang tepat penting untuk penyelidikan ternakan. Kaedah tradisional untuk merekodkan berat badan ternakan menggunakan alat konvensional kurang tepat. Penerapan kaedah analisis imej untuk penentuan berat badan adalah satu teknik yang sesuai dan tepat memandangkan pengukuran dimensi imej haiwan adalah secara automatik dan menggunakan analisis prediktif.

Pengesanan mastitis

Mastitis adalah penyakit utama lembu tenusu yang menjejaskan kesihatan haiwan ternakan dan mengganggu pengeluaran susu. Data pemerah susu seperti kekonduksian elektrik, jumlah susu, data enzim laktat dehidrogenase dan skor sel somatik biasanya diperoleh dari masa ke masa melalui mesin pemerah susu automatik. Data-data ini berserta

dengan data ujian makmal berkala dan data ujian diagnostik doktor dapat digunakan untuk pengesanan mastitis dengan menggunakan kaedah rangkaian neural. Penggunaan kaedah rangkaian neural dalam mengenal pasti jangkitan awal mastitis membawa kepada pemulihan ternakan di kawasan ladang apabila langkah-langkah pencegahan diambil di dalam kawasan kritikal yang telah dikenal pasti.

Cabaran analisis data besar

Analisis data besar membawa banyak peluang menarik terutama dalam menyelesaikan masalah-masalah masa-nyata terutama berkaitan dengan industri ternakan. Namun begitu, terdapat banyak cabaran yang dihadapi ketika menangani masalah data besar. Cabaran utama terletak pada kaedah penangkapan data, penyimpanan, pencarian, perkongsian, analisis dan isualisasi. Sekiranya cabaran ini tidak dapat diatasi, data besar ini akan hanya menjadi sia-sia sahaja kerana kita tidak mempunyai keupayaan untuk menerokanya.

Ketidakseimbangan sistem pemprosesan masih mengekang perkembangan penemuan data besar. Mengikut Hukum Moore, prestasi unit pemprosesan berpusat [*Central processing unit* (CPU)] meningkat dua kali ganda setiap 18 bulan dan prestasi pemacu cakera juga meningkat dua kali pada kadar yang sama. Walau bagaimanapun, kelajuan putaran cakera hanya bertambah sedikit sahaja sejak sedekad yang lalu. Akibat ketidakseimbangan ini, kelajuan input/output (I/O) rawak telah meningkat sedikit sahaja manakala kelajuan I/O sekuensial hanya bertambah dengan kadar yang perlahan. Selain itu, maklumat semakin meningkat pada kadar eksponen, tetapi teknologi dalam pemprosesan maklumat membangun secara perlahan.

Dalam banyak aplikasi data besar, teknik dan teknologi yang canggih tidak dapat menyelesaikan masalah sebenar, terutamanya untuk analisis masa nyata. Sehingga kini, saintis dan juruanalisis data masih tidak mempunyai alat yang tepat untuk mengeksploitasi data-data besar ini. Cabaran dalam analisis data besar antaranya termasuk data yang tidak konsisten, data yang tidak lengkap, ketepatan masa dan keselamatan data. Memandangkan saiz set data seringkali sangat besar sehingga mencapai terabait, pangkalan data yang ada seringkali menghadapi masalah mengenai data yang tidak konsisten, tidak lengkap dan hingar. Oleh itu, beberapa teknik prapemprosesan data, termasuk pembersihan data, integrasi data dan transformasi data boleh digunakan untuk menghilangkan data hingar dan memperbaiki ketidakkonsistenan data.

Kesimpulan

Analisis data besar masih berada di peringkat awal pembangunan, kerana teknik dan perkakasan yang ada untuk analisis data besar sangat terbatas dalam menyelesaikan masalah data besar sepenuhnya. Daripada perkakasan ke perisian, analisis data besar memerlukan lebih banyak masa dan usaha ditumpukan dalam pembangunan kaedah penstoran data yang lebih canggih, arkitektur komputer yang lebih baik dan teknik intensif data yang lebih efisien (sebagai contoh: pengkomputeran awan, pengkomputeran grid dan pengkomputeran biologi) dan teknologi yang lebih progresif. Selain itu, komuniti sains ternakan hari ini tidak mempunyai infrastruktur dan alat untuk dimanfaatkan sepenuhnya dalam memproses lambakan data-data yang diperoleh. Apabila data-data molekular seperti genomik, transkriptik dan mikrobiota pada haiwan berjaya diperoleh, teknik analisis data besar seperti teknik perlombongan data, pembelajaran mesin dan analisis rangkaian neural boleh mengekstrak maklumat penting serta memberi rumusan dan menyelesaikan permasalahan dalam industri ternakan. Kerjasama yang rapat antara pelbagai disiplin ilmu seperti sains komputer, ekonomi, kejuruteraan, matematik dan statistik, bersama-sama dengan pemacu industri ternakan amat diperlukan supaya teknologi ini dapat diperkembangkan lagi.

Bibliografi

Boyd, D. dan Crawford, K. (2012). Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication, and Society* 15(5): 662 – 679

Gomes, R.A., Monteiro, G.R., Assis, G.J.F., Busato, K.C., Ladeira, M.M. dan Chizzotti, M.L. (2016). Estimating body weight and body composition of beef cattle through digital image analysis. *J. Anim. Sci*. 94: 5,414 – 5,422

González-Recio, O., Rosa, G.J.M. dan Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livest. Sci.* 166: 217 – 231

Long, N., Gianola, D., Rosa, G.J.M., Weigel, K.A. dan Avendano, S. (2007). Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *J. Anim. Breed. Genet*. 124: 377 – 389

Morota, G. dan Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. *Front. Genet*. 5: 363

Philip Chen, C.L. dan Chun-Yang, Z. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences* 275: 314 – 347

Sun, Z., Samarasinghe, S. dan Jago, J. (2010). Detection of mastitis and its stage of progression by automatic milking systems using artificial neural networks. *J. Dairy. Res*. 77: 168 – 175

Ringkasan
Pertumbuhan pesat data ternakan daripada alatan sensor, teknik penyelidikan terkini dan daripada media elektronik telah membawa kepada peluang dan cabaran baharu dalam industri pertanian. Lambakan data yang diperoleh di ladang ternakan dan makmal sains tidak dapat dimanfaatkan sepenuhnya kecuali ia dianalisis dengan tepat untuk menghasilkan maklumat kritikal yang kemudiannya boleh digunakan untuk membuat keputusan. Alatan pertanian tepat bersama-sama dengan prosedur perlombongan data dan kerangka pembelajaran mesin, dapat mengekstrak maklumat kritikal daripada data besar yang diperoleh. Artikel ini bertujuan menerangkan tentang aplikasi, peluang dan cabaran data besar dan bagaimana analisis data besar dapat digunakan untuk menyelesaikan masalah dalam industri ternakan.

Summary
The exponential growth of livestock data from real time sensors, new breakthrough techniques and from other electronic media has led to new opportunities and challenges in agricultural industries. The vast amount of data being generated on livestock farms and animal science labarotary is meaningless unless it is analysed in time to yield critical information which can then be employed to make a decisions. Precision agriculture tools, when coupled with advanced data-mining procedures and machine learning framework, can extract critical information from big data. This article is aimed to demonstrate a close-up view of the applications, opportunities and challenges of big data, and a glimpse into how they can be applied to solve pressing problems in livestock industries.

Pengarang
Mohd Azri Azman
Pusat Penyelidikan Sains Ternakan, Ibu Pejabat MARDI,
Persiaran MARDI-UPM, 43400 Serdang Selangor
E-mel: mohdazri@mardi.gov.my